The emergence of norms in society

Lisciandra, C.

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

The Emergence of Norms in Society: A Philosophical Investigation

Chiara Lisciandra

The Emergence of Norms in Society: A Philosophical Investigation

Chiara Lisciandra

The Emergence of Norms in Society: A Philosophical Investigation

Proefschrift
ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 25 oktober 2013 om 14.15 uur
door

Chiara Lisciandra

geboren op 31 juanari 1983 te Milaan, Italië

Promotor:	Prof. dr. S. Hartmann
Copromotor:	Dr. J.M. Sprenger
Overige Leden:	Prof. dr. J. Alexander
	Prof. dr. J.W. Romeijn
	Prof. dr. A.P. Thomas
	Prof. dr. J. Vromen

Pieter Brugel the Elder (1525-1569), *The Procession to Calvary,* 1564.

Acknowledgments

Philosophy has been my loyal and omnipresent companion during the years of my PhD. Luckily enough, I had the pleasure to share its company with all the people with whom I have worked and who made this relationship so intense and lively.

First of all, I would like to thank my supervisor, Stephan, for his support during my studies in Tilburg and in Munich. I owe him thanks for having encouraged me to travel around the academic world and because it has always been such a pleasure to work with him.

A sincere thanks to my copromotor, Jan, who has actively and closely participated in the final stage of my PhD.

I am extremely grateful to Jason Alexander, Jan-Willem Romeijn, Alan Thomas, and Jack Vromen for having agreed to be part of my committee thesis and for their invaluable comments.

Also, a special thanks goes to my previous mentor, Francesco Guala, who first directed me towards this wonderful philosophical experience.

My 'scientific siblings' have an important role in all my academic achievements. Thanks to Marie Postma, Carlo Martini, Matteo Colombo, and Ryan Muldoon. With them, I have had the opportunity to discuss and continuously confront my ideas and to establish philosophical affinities. I consider myself extremely fortunate to have had such excellent co-authors.

During my detours in a foreign language, Anne has guided me with patience through the English writing, its literature and sounds. Thank you for that! Last but not least, I would like to thank two great secretaries, Annette and Nicole. I wish I could have taken them with me to any university I will be at in the future

For the life besides my academic work, I don't even know where to start. During the PhD, I have travelled across different countries and had to realize how it feels to belong to very distant realities: somehow, very rich and very

poor at the same time. Today, I find myself with my heart in Italy, my head in The Netherlands, my cold blood in Helsinki, my dreams in the US, and my thoughts in Germany. In each place I have people to thank: my friends in Milan for being a constant certainty and for helping me to maintain our friendship despite the distance. Thanks to my lovely guys in Tilburg, who make me want to stay here every time I come back. Also, thanks to the MCMP crew, from whom I learned how to combine a lot of work and a lot of fun!

I cannot name all the people to whom I would like to dedicate this work. But let me mention two persons who are behind all I have done so far. My greatest thanks are to my parents, for being so happy about the chance I had in life. Papà, thanks for reading patiently all the versions of my papers! And, Mamma, thanks for being my best confidante ever and for being patient even if the distance of these years has been tough. I hope the possibility will come to make my way back home.

Grazie a tutti! Chiara

Abstract

My research is on the role of identity and norms in economic decision making. In particular, I focus on the emergence of norms and study how behaviors which were not originally regulated by norms gradually become entrenched practices and acquire a normative force. For this purpose, I develop probabilistic models which help illustrate the features of the emergence of norms in society. This research is accompanied by a family of experimental studies on the effects of social cues on norms compliance. In my research, I use a combination of formal and empirical methods and explore the conditions that make, or do not make, formal models an appropriate tool for describing social phenomena and suggesting interventions in society. These and related topics have been elaborated in my doctoral thesis, which is an interdisciplinary work combining different approaches and methodologies to investigate norm-related behaviors. Each chapter constitutes an autonomous scientific paper which addresses a different question. The first chapter explores a series of probabilistic models for the emergence of descriptive norms in society. The second provides an explanatory framework for descriptive norms, according to which they originate as a by-product of a Bayesian updating process for detecting regularities in the natural world. The third chapter offers an experimental study to delineate a novel taxonomy of normative judgments on the basis of their insulation from group conditioning. Finally, the fourth chapter provides a methodological reflection on robustness analysis, as a non-empirical confirmatory tool employed in scientific practice.

Contents

We hereby report you
The story of a journey, undertaken by
One who exploits and two who are exploited
Observe the conduct of these people closely:
Find it estranging even if not very strange
Hard to explain even if it is the costum
Hard to understand even if it is the rule
Observe the smallest action, seeming simple
With mistrust
Inquire if a thing be necessary
Especially if is common
We particularly ask you
When a thing continually occurs
Not on that account to find it natural
Let nothing be called natural
In an age of bloody confusion
Ordered disorder, planned caprice,
And dehumanized humanity, lest all things
Be held unalterable.

Bertolt Brecht. The Exception and the Rule.

Chapter 1

Introduction

The norms we live by are a dynamic entity in our societies. Norms continuously emerge, develop and establish themselves in groups. At times they are dismissed, only to be adopted again before being substituted with new ones. Just as with language, the norms of society characterize our culture. When we act in accordance with them, we communicate who we are and where we come from.

This study focuses on a distinct set of norms, which are those that emerge spontaneously from repeated interactions between individuals of the same group. Each society is filled with myriad norms of this kind. Examples are conventions, moral norms, social and descriptive norms. Norms of fairness, trust, greetings codes, dress codes and etiquette, are cases in point. Even if they are not part of a codified system, these norms distinguish what is allowed from what is not within a social group. Even if we do not see them, they regulate many small features of our interactions. Informal norms, together with those norms that will eventually become part of a written code, contribute to the construction of our social reality.

Consider this scenario by David Lewis (1969):

> *Suppose that with practice we could adopt any language in some wide range. It matters comparatively little to anyone (in the long run) what language he adopts, so long as he and those around him adopt the same language and can communicate easily. Each must choose what language to adopt according to his expectations about his neighbors' language: English among English speakers, Welsh among Welsh speakers, Esperanto among Esperanto speakers, and so on. (pp. 7-8)*

The philosophical puzzle arising from Lewis's description is how to decide which language to adopt in the first place. In principle, I will adopt English if I expect that the other speaker will adopt English, who in turn will adopt English if he expects that I will adopt English, which again depends on whether he expects that I expect that he will adopt English, and so on.

The situation with norms is similar. Once established, the norms of society convey signals by virtue of mutual agreements between members of the same group. In this sense, when we interpret the actions of other people as abiding by or violating norms, we judge them with respect to a set of mutual expectations. Just as with words, we can name a *house*, a *casa*, or a *maison*, so in the case of norms it does not matter whether we decide for X or Y, for example whether to greet each other with handshakes, bows, or kisses, as long as everybody does the same. In this respect, these norms are arbitrary in the same way as the words of language are. The question of how certain norms emerge, when other possible norms are on an equal footing, will be addressed in the first part of this thesis.

Lewis's analysis (1969) focused on the notion of social conventions, for the study of which he employed the conceptual apparatus of rational choice theory, and was largely influenced by the previous work of Thomas Schelling (1960). One of the most significant contributions of Lewis and Schelling to the field was to show the potentials and limitations of game theory when applied to social ontology. As both authors made clear, game theory is a powerful tool with which to represent interactive decision problems. But game theory alone is inadequate, if not supported by an analysis of the actual mechanisms regulating individuals' decision-making in interactive contexts.

As a response to this critique, an interdisciplinary research program has been started which combines rational choice theory with the study of decision-making processes coming from cognitive and social psychology. The analysis of conventions has since been extended to other kinds of norms, i.e. social norms, moral norms, descriptive norms, etc. In this area of research, the criteria for differentiating between different kinds of norms are still a matter of dispute and the second part of this thesis will present an experimental study that advances a new criterion for norm taxonomy.

Overall, it is the purpose of this thesis to contribute to a research agenda whose main motivation is to remedy the several flaws of a theoretical approach to interactive decision-making, which sees the *homo economicus* as the standard against which to measure the canons of human rationality. In order

to avoid the shortcomings of a rational choice approach to the study of norm compliance, the studies presented in this work attempt to combine in a unified picture an analysis of human cognition and a theory of action underlying individual decision-making in group contexts.

To this end, in the course of my PhD I have explored the processes that lead to the emergence of norms and their compliance using different methodologies, i.e. formal models, simulations, and experiments. When giving talks to present my work, I have often been asked to clarify how this study differs from related works in sociology, psychology or economics. In other words, why is this topic the subject of a philosophical investigation? On my side, I have found solid philosophical support. I could answer the question in the way Popper did in *The Logic of Scientific Discovery*: *"I do not care what methods a philosopher (or anybody else) may use so long as he has an interesting problem, and so long as he is sincerely trying to solve it."* (Popper 1959, p. XX). Additionally, I consider this to be a subject of a genuine philosophical nature. As will appear from the following pages, the questions I address in this thesis combine two aspects which constitute a philosophical inquiry, namely ethical analysis and scientific research.

For one thing, the process that leads to the emergence of norms is regulated by inner mechanisms that need to be elucidated. The norms I am interested in are not the result of the decisions of authorities or policy makers, but are the outcome of unplanned, bottom-up processes with their own evolutionary paths. For another, unveiling the conditions behind the adoption of new norms is of primary relevance to actual society. A deeper understanding of the dynamics of change in norms could facilitate the integration process between cultures with conflicting sets of norms. Furthermore, it could indicate how to accelerate the decay of inefficient or negative norms, such as discriminatory norms, unhealthy conduct, or unsustainable behaviors and, at the same time, guide the introduction of positive ones, for example environmental or public-health policies.

Overall, this work attempts to clarify controversial philosophical issues related to the emergence of norms and norm compliance. However, I acknowledge that the methods I adopt for this analysis are not mainstream in philosophy, as they are not restricted to conceptual analysis, thought experiments, and case studies. In this respect, I value the increasing tendency towards a scientific approach to philosophy, which I also endorse by employing formal and experimental methods for the solution of philosophical

problems.

As will appear, the questions I address are apt to be considered within a broader methodological framework rather than a purely analytical one. At the same time, translating conceptual problems into models and experiments opens a number of methodological issues that are as crucial as the questions for which these tools were originally adopted. With these issues in mind, the third part of this thesis will be dedicated to a methodological reflection on robustness analysis, which is a method of comparing the results of a set of several experiments and/or theoretical models that have been formulated to investigate the same class of phenomena.

This thesis consists of three parts. In the first part, I focus on a specific set of norms, namely descriptive norms. Classical examples of the sort are fashion, fads and trends. In the second part, I investigate the distinction between different kinds of norms, i.e. moral, social and decency norms. In the third part, I consider some methodological questions related to robustness analysis. The subject matter of each section can be expressed in the form of a question:

- How do descriptive norms emerge?
- How do we selectively distinguish between moral, social and decency norms?
- What are the assumptions underlying robustness analysis, as a method of non-empirical confirmation of scientific theories?

In the remaining part of this introduction, I will briefly present the content of each section in the methodological context in which it is explored.

Topic and Methods

The Formal Approach: Two Studies on the Emergence of Norms

The study of social phenomena through models and simulations has its origin in the seminal work of Thomas Schelling on racial segregation (1971). Schelling studied how macro-phenomena can emerge as the unintended effect of the combination of many individual decisions. The novelty of this approach lies in the fact that Schelling investigated social phenomena not

by looking directly at their macro-variables, but explained them as the unplanned consequences of aggregated individual interactions. Racial sorting is a case in point. Schelling's model shows that segregation occurs not as a consequence of the preferences of the individuals for segregation itself, but as the indirect effect of the preference of individuals for having a few neighbors of the same ethnic group.

Schelling represented the segregation process by means of a checkerboard, with dimes and pennies, standing respectively for a certain metropolitan area and for the individuals of two different groups. This is why Schelling's model is an example of a *paper and pencil* model. On the checkerboard, it is possible to trace the movements of the individuals, and to observe how the configuration of the neighborhood changes as a consequence of the individuals' decisions to move to areas where they will have some neighbors of the same group.

If the same model were implemented in a computer simulation, then it would be an instance of an agent-based model.[1] Agent-based models, such as those I have developed in the second and third chapter of this work, are a class of computational models that study the dynamic of interactive systems. By relying extensively on computer simulations, agent-based models considerably increase the predictive power of traditional models: they make it possible to analyze phenomena involving a large number of factors and their aggregated effect, thereby overcoming the problem of analytic tractability of non-simulated models. Through them, we can formulate hypotheses based on fewer idealizations and whose degree of proximity to the target system is higher than it would be without such simulating devices. Because of their computational power, agent-based models have become an indispensable research tool across disciplines, and nowadays they are not exclusively employed in the social and behavioral sciences, but also in economics, biology, ecology, epidemiology, etc.

As an illustration, let us consider a simplified reconstruction of an agent-based model for a social phenomenon. The premise is to build a model which captures the relevant variables of the individual decisions, such as personal preferences, response to other agents' behavior, and to the context. Next, a way has to be found to implement the model and the other components that

[1]Notice, however, that there is not a difference in principle between paper-and-pencil models and computational models: they are both instantiations of a universal Turing machine, even if usually only the latter require extensive computational resources.

characterize the system – such as the network structure – into a computer code. Call the set of relevant factors that are external to the model its *environment*. Together, the model and the environment constitute the algorithm which runs on the computer. Each run of the program corresponds to a step of the simulation, which in turns represents a change in the system. The evolution of the system can be represented graphically by means of software that transforms the numerical analysis into visual representations.

The procedure outlined above is the one I have followed to build two agent-based models for the emergence of norms. Whereas Schelling's model describes segregation as the consequence of the preference of individuals for having a percentage of similar neighbors, I consider how the preferences of the individuals and their interactions are conducive to a norm's emergence. More specifically, I am interested in how individual decisions, which result from a combination of personal preferences, expectations about other people's behavior, and social influence, aggregate and give rise to norms. To do so, I have been working on a set of probabilistic models to answer the question: how do behaviors, which were not originally regulated by norms, gradually become entrenched practices and acquire a normative force?[2]

In the first study, I represent the individual decision to follow a norm as the outcome of an heuristic process; in the second, as the outcome of a Bayesian deliberation. In both cases, I focus on the emergence of a specific set of norms, namely descriptive norms. These are rules of behaviors mainly driven by a combination of individual attitudes and the desire to conform. They involve only one level of expectations, i.e. what we expect other people will do in similar circumstances. Other norms, such as social norms, also involve beliefs concerning what we think other people expect us to do. The study of descriptive norms constitutes the starting point of an analysis that in the future will be extended to norms that involve higher level expectations.

In the second chapter of this thesis, my co-authors and I use a standing

[2]The literature on decision-making processes mediated by norms, conventions and the like has –as its starting point– the foundation of morality and the central is-ought-question. However, in its development it has also radically departed from the meta-ethics aspects of the problem. In the rest of the text, when we talk about normativity, we refer to normative expectations, namely the beliefs about what other people expect us to do. When we talk about certain behaviors acquiring normative force, we refer to the set of mutual expectations that progressively become regulative in interactive decision-making processes. Overall, the purpose of this area of enquiry is to analyze how certain aspects, such as the context or the framing, affect our perceptions of what is normative and influence our choices accordingly.

ovation effect as a metaphor for the emergence of descriptive norms. Standing ovations often occur after theater plays or sports competitions, when the spectators in the audience progressively stand up to express their appreciation of the performance. The rationale behind the modeling trick is that descriptive norms resemble the character of a standing ovation, insofar as they are both contagious effects deriving from idiosyncratic preferences and group influence. As the second chapter shows, with just a few modifications of the baseline model, it is possible to apply the standing ovation analysis to the case of descriptive norms.

Whereas the first model presents a cognitively realistic mechanism for norms emergence, the second provides a rational reconstruction of individual decision-making. More specifically, in the case of the Bayesian model, the members of a group act as Bayesian updaters: they start with certain priors about the desirability of a norm, which they revise after having considered the external evidence, i.e. whether or not other agents follow the same behavior. The reliability of other individuals is calculated by the likelihood ratio, exactly the same way as, for example, the reliability of a diagnostic test is calculated in clinical trials. Individuals formulate the posteriors; and the higher the posteriors are, the higher the probability that they will follow the norm.

This way of proceeding shows how the basic reasoning process of Bayesian updating can be transformed, with just a few modifications of its mathematical machinery, into a mechanism of social rule discovery. The success of Bayesian belief revision in dealing with the natural world provides a reason why we might find individuals naturally extending this apparatus to the social world. When this decisional problem is implemented in an agent-based model, we can observe the conditions under which a new norm becomes established in a group, as more and more people comply with it, and the speed of convergence on that norm. The results of the Bayesian model can then be compared with those of the heuristic one, and the robustness of their predictions analyzed.

The foregoing examples illustrate what I take to be the proper method of approaching this kind of research. The starting point is the translation of a decisional rule into a mathematical model, whose predictions can be observed by means of formal analysis and computer simulations. The model's predictions can be tested afterwards by means of laboratory experiments, which in turn can provide feedback about the normative model.

The Experimental Approach: A Study on Group Conditioning of Normative Judgment

The second part of this thesis presents the results of an experimental study on norm compliance. The topic and the methods employed make it an example of a work in experimental philosophy, a field of study which sees philosophical investigation directly involved in empirical research. Traditionally, the empirical sciences have been the subject of philosophical study, albeit mainly in an *a priori* fashion. Philosophers of science look at a variety of aspects of science, i.e at notions such as explanation, prediction and causation; or at different criteria of theory confirmation, the validity of specific research methods and the results to which they are conducive. This sort of analysis, however, constitutes a theoretical approach to the scientific enterprise.

In tandem with the traditional approach, recently there has been a philosophical shift in attitudes to experimental research, according to which observational evidence should constitute the raw material that informs philosophical knowledge. This tendency has been of interest to philosophers working in several different areas of their own discipline, from the philosophy of language to the philosophy of psychology, to logic, decision theory etc. My research has focused specifically on moral psychology, a field of study which addresses questions in ethics, not only via conceptual analysis, but also by evaluating individual responses to morally-loaded situations. Broadly speaking, moral psychology explores questions such as whether we can talk about a 'moral sense', and if so, how it evolved; it investigates and compares people's normative intuitions, asks how automatic emotional reactions combine with higher-level information related to normative reasoning, and seeks the neural correlates of normative judgments (Greene et al. 2004; Haidt 2001; Prinz 2006). To address these and related issues, data are gathered from behavioral, cognitive and neuroscientific studies.

Within this methodological framework, I have carried out a family of experiments, together with an ethicist and a statistician based at Tilburg University, to collect behavioral data on the criteria that individuals adopt to distinguish between different kinds of normative judgments.

In ethics, normative judgments are those related to 'ought to' statements. From a purely theoretical point of view, if normative judgments express what 'ought to be done', it is immaterial whether the content of the prescriptions refers to moral norms, involving principles of justice and welfare, or to social

norms, involving principles of fairness which might change across contexts. However, it is also evident that we judge differently the violation of a norm, as for example stealing, and that of a dress code, such as wearing pink at a funeral. In the literature of normative judgments, the criteria for distinguishing between different kinds of norms have been investigated, but the existing accounts have so far failed to identify a consistent taxonomy (Kelly et al. 2007; Nichols 2002; Sousa et al. 2009; Turiel 1983).

The study presented in the fourth chapter of this thesis investigates whether a novel classification of norms can be based on their degree of dependence on other people's beliefs and preferences. In other words, we ask whether a distinction can be drawn between different kinds of normative judgments on the basis of their sensitivity to peers' opinions.. Overall, the aim of this work is to make progress both in understanding which features allow our minds to distinguish selectively between different kinds of norms, and more specifically how social cues impact normative judgments.

The experimental setup we adopted is a modification of the Asch paradigm (1951, 1955). Asch examined the effects of conformity in a group of subjects exposed to a visual perceptual task. The task simply consisted of matching different lines according to their length. Despite the easiness of the task, Asch showed that when the experimental subjects were in a group of confederates all giving the same wrong answer, they tended to align to the uniform wrong answer. In our study, we replicate the Asch experiment in the moral domain. A pool of experimental subjects were recruited to take part into an experimental setup consisting of two parts: first, subjects were asked to fill in anonymously a survey with several short scenarios representing the violation of different kinds of norms, namely moral, social and decency norms. Following this, the same subjects were called into a university office and asked to evaluate the same scenarios in the presence of a group of confederates. Unbeknown to the real subjects, the confederates were previously trained to provide different answers from those given by the subjects in the individual questionnaire. The experimental findings will be presented at length in the fourth chapter. In a nutshell, they indicate that the degree of conformity differs according to the type of norms at issue. Interestingly, the results showed that moral norms are subject to conformity, especially in situations with a high degree of social presence.

On Robustness Analysis

The last chapter of this thesis presents a methodological reflection on robustness analysis, which is a method of studing a certain class of phenomena through different experimental or theoretical methods, and discusses its philosophical significance. The idea behind robustness analysis can be illustrated by means of a simple example. Suppose you had two different watches and that, just before leaving to go to a meeting, you checked the time on both of them. If the time displayed is the same, then the confidence that the watches are functioning properly is higher than if they were not, in which case other evidence should be taken into account. Analogously, robustness analysis in science is a method of ascertaining the accuracy of result via multi-modal verification.

In the experimental sciences, robustness analysis is a method of testing the effect of possible confounders on empirical results. Given that experimental settings are simplified representations of real-world situations, then it is a necessary preliminary to check whether their results do not depend on one of these simplifications. Similarly, robustness analysis in the non-experimental sciences is a method of testing whether the predictions of the model are the unintended effect of certain specific theoretical assumptions. Given that scientific models are based on abstractions and idealizations, which do not literally mirror the target system, a test of robustness proceeds by changing some of the unrealistic assumptions to check whether the same result holds true across conditions.

Intuitively, robustness analysis strikes one as a plausible method of inquiry. Still, a number of skeptical arguments have been raised against its confirmatory value. The aim of the final chapter of this thesis is to present the main positions of the debate and to investigate the assumptions behind robustness analysis in theoretical modeling.The motivation behind this study stems from the fact that an appeal to robustness will repeatedly be made throughout this thesis. Different models for the emergence of norms will be presented and the importance of their robustness highlighted. Similarly, the results of a family of experiments on norms compliance will be discussed, where the idea underlying the variations in the setup is that of adding progressively to an ordered sequence of investigations. Overall, robustness analysis is inherent in the process of model building and designing experiments. Considerable literature has been published on experimental ro-

bustness (Stegenga 2009; Soler 2012), there is a lively ongoing debate on the status of theoretical robustness analysis (Kuorikoski et al. 2010; Odenbaugh and Alexandrova 2011; Woodward 2006). In an attempt to clarify the notion, a general distinction will be explored between robustness analysis as a method of testing the role of different assumptions about the system being investigated, and as a method of testing the role of the assumptions introduced into a model for reasons of mathematical tractability. I will defend the claim that the comparison of the results coming from models based on different assumptions is in principle helpful, but that the method of conducting this sort of analysis is far from straightforward.

As I will explain in more detail, robustness analysis has become an umbrella term that covers different meanings. Drawing on a distinction suggested by Weisberg and Reisman (2008), I refer to *parameter*, *structural* or *representational* robustness, depending on the object subjected to variation in the model. In accordance with this classification, a test of robustness is made in the second chapter by modifying the parameters of the model and its structure. There, we start from a baseline model, i.e. a very simplified representation of the phenomenon under study, and progressively add more and more realistic elements to its structure. In the third chapter, we address the problem of the emergence of norms within a formal epistemology framework, that helps to motivate the assumptions made in heuristic models. The idea behind an heuristic and a rational model is that of focusing on different aspects of the problem under consideration. Whereas an heuristic model can take into account the psychological mechanisms of norm compliance, a domain-general model of belief revision helps explain the disposition to look for the regularities that generate descriptive norms at the outset.

Despite the advantages of robustness analysis in the studies presented above, I will argue that opportunities for conducting this practice are often limited. This means that robustness analysis is by necessity confined to models whose structure is relatively simple and whose predictions can been compared with one another. As I will explain throughout the chapter, however, this is not always the case in science, especially with models with a particularly complex structure, as for example in economics. I will indicate some of the difficulties encountered in the practice and suggest possible alternatives, focusing on a case study in economic geography.

Chapter 2

On the emergence of descriptive norms

Descriptive norms hide in plain sight.[1] While we may not always think of them, they govern many of our day-to-day interactions: they help guide our fashion choices, our table manners, the colors we wear at weddings, and any number of other small features of our social interactions. This governance can become evident when we travel: many of our small-scale behaviors and interactions are culturally contingent. Americans typically greet each other with handshakes. Many continental Europeans greet each other with kisses to the cheek – but the number varies between countries. Standards for personal space vary across cultures. It would be difficult to argue that any one of these practices is 'right' – descriptive norms do not carry the normative weight of social or moral norms – but we all follow the norms from our own cultural context, and imitate the behavior of our peers. While this may provide a satisfactory account of how descriptive norms operate, it does not tell us about how they came to be. What is it about a given norm that caused everyone to start following it?

To begin to answer this question, we will turn to a simple case, that of the standing ovation. Standing ovations, like many other descriptive norms, are the result of spontaneous coordination of individual choices across many individuals. They have become a common practice after many live performances, even though there is no pre-arranged plan or even any formal coordination. All individuals can do is decide whether or not they wish to stand, based on

[1] This chapter is based on Muldoon, Lisciandra, Bicchieri, Hartmann and Sprenger (forthcoming).

their own preferences, and what they see others doing around them (Miller and Page 2004). Put slightly more formally, most agents have preferences about whether or not they like to stand up, which depend on the quality of the concert. They know that standing to clap is a common option after a performance, and they have (conditional) preferences for standing up if the other agents stand up, too. This exemplifies a *descriptive norm*: the agents know that there is a behavioral pattern (standing to clap) that applies to the situation they are in, and they prefer to conform with the group (Bicchieri 2006, 31-32). In other words, their behavior is not only determined by unconditional preferences for certain actions, but also by their desire to conditionally conform to the behavior of a sufficiently large group.

Standing ovations are a useful stand-in for describing societal transitions to a wide variety of descriptive norms. That is, there is a status quo behavior that can be upended by an alternative. For example, fashion often works in this way. Prior to mini skirts, women wore longer-length skirts, and upon their introduction, the population largely shifted to a mini-skirt norm. Movements in popular music also follow a similar pattern: teenagers largely coalesce around a few bands or a particular kind of music for a few years, before giving way to a new set of music. Calling etiquette has similarly shifted as email has become a more common form of communication. What we find in all of these cases is a status quo that, without any central coordination, loses out to a new behavioral rule. By focusing on a standing ovation model as an exemplar of this wider set of phenomena, we can avoid the problem of getting lost in small case-specific details, and instead try and identify the key features of individual decision-making that can affect the emergence of a norm. Thus, we aim to examine how ovations might arise, and in doing so, come to a more general account of the emergence of descriptive norms in a population.

To provide for this more general account, we investigate several features of individual decisions, such as a desire to conform, one's knowledge of what others are doing, and one's own preferences. These elements can affect the emergence of a descriptive norm in a group and influence some aspects of the processes, such as whether the group converges on a single behavior, and if it does, how quickly this happens. Our model allows us to carefully explore the key aspects of individual decision-making that drive these collective behaviors.

One additional fact that we wanted to take into consideration was that

though descriptive norms can be built out of many small decisions, they do not always emerge. While many of our day-to-day activities are governed by norms, not all of them are. Plenty of our actions are different. We do not all walk in lock-step. Further, many of our seemingly coordinated actions can be simply described as behavioral regularities – agents act purely in accordance with their intrinsic preferences, which just happen to align with others'. Descriptive norms, on the other hand, arise when the desire to conform to the behavior of others overwhelms one's initial preferences. Our model helps us to explore the contingent nature of many descriptive norms. What this model suggests is that it is possible that some descriptive norms become descriptive norms for no particular reason other than the peculiarities of the individuals in the population.

In this study we explore four main models:

1. In the first model, *the baseline*, we build a model for a standing ovation, which considers an individual's decision about whether she will stand to be a combination of her personal unconditional preference and her tendencies to choose to conform to the behavior of others.

2. In the second model, *the inertia model*, we introduce two new features of individual decision-making: first, a tendency of individuals to become increasingly set in their ways as time goes on, and second, a more nuanced model of social contagion, to better match how bandwagon effects occur.

3. In the third model, *the endogenous social sensitivity model*, we treat one's sensitivity to the behavior of others not as something separate from one's individual preferences, but as dependent on them.

4. In the fourth model, *the symmetric model*, we consider a reversible case, where each agent can decide whether to stand up or to sit down in any round. We assume that the population is made by two agent-types with opposite preferences. This setting allows us to compare the emergence of descriptive norms, where all individuals set on the same action regardless of their own preferences, and behavioral regularities, where all individuals continue to follow their intrinsic preferences despite other peoples' actions.

Each model helps us learn more about the nature of social decision-making. The first three models explore the robustness of the standing ovation app-

roach as an explanation of directed norm emergence. These three models together allow a more nuanced view of social effects on individual decision making than just the baseline standing ovation model would allow. By examining more decision procedures, we are able to better describe a larger class of directed norm transitions than we would otherwise be able to. Being able to look at information asymmetries between agents further allows us to examine the effects of social hierarchies on norm emergence. The fourth model that we consider widens the scope of our investigation into norm emergence by examining a symmetric case that allows for norm emergence in either direction, or none at all. Here we are able to more fully investigate the conditions for when norms do not emerge, since the model allows us to compare behavioral regularities from descriptive norms, as we can directly inspect agent preferences. Unlike Lewis's approach to the analysis of conventions, according to which conventions are solutions to coordination games where 'each wants to conform conditionally upon conformity by others' (Lewis 1969), we focus on the decisions of agents who do not reason about other agents' expectations, and on norms that themselves have no intrinsic coordination advantages. We are interested in studying the dynamics created by individuals who have both intrinsic preferences to act and some interest in conformity. This allows us to focus on a wider range of more common descriptive norms, many of which end up having large cultural variation. While much of this analysis could apply to Lewis-conventions, we examine a less structured environment. In the family of models we consider, not only do we allow for agents to have interests beyond pure coordination, we allow for asymmetries between agents across several dimensions, and consider both complete and incomplete information conditions.

Together, these models provide a more general account of descriptive norm emergence than has been seen so far in the existing literature. First, by focusing on norms that have no inherent differences in utility, we can focus on the large class of under-studied everyday norms, such as norms of personal space, etiquette, dress, eye contact, and other small-bore issues. These norms, when put together, help explain a large portion of our social behavior, even if any individual norm has only a small effect. Second, by introducing information asymmetries between agents, we begin an analysis of the effects of social hierarchies on norm dynamics. Third, by splitting our study of norms into the directed and bidirectional cases, we can then study the differences between behavioral regularities and descriptive norms, which

can otherwise get lumped together in other literature.

2.1 The Baseline Model

2.1.1 Model Description

Let there be M people in the audience of a theater play. The variable $s_i^{(n)}$ with $i = 1, \ldots, M$ indicates whether person i is sitting ($s_i^{(n)} = 0$) or standing ($s_i^{(n)} = 1$) at time-step ('round') $n = 0, 1, \ldots$. Time is discretized and at $n = 0$, everybody is seated. Everybody who is not yet standing 'updates' her position in each round. Our central idea is that whether or not a person stands up depends on her *effective propensity* to do so. The effective propensity of person i is the convex combination of two factors:

1. An *intrinsic preference* q_i to stand up. This represents an individual's preference to stand up or not, independently of what other people are doing.

2. An *extrinsic propensity* to stand up. This factor takes into account what other people in the audience are doing. So whether or not someone stands up in round n will depend on how many people $S^{(n-1)} := \sum_{i=1}^{M} s_i^{(n-1)}$ were standing up at round $n-1$. Note that $S^{(0)} = 0$.

Combining these two factors, we arrive at the following expression for the *effective propensity* to stand up in round n:

$$P_i^{(n)} = \sigma_i \left(\frac{S^{(n-1)}}{M} \right) + (1 - \sigma_i)\, q_i \tag{2.1}$$

We see that the variable $\sigma_i \in (0, 1)$ measures how much person i takes the extrinsic propensity, i.e. social considerations into account. It determines the relative weight of intrinsic and extrinsic propensity. Therefore, we say that σ_i measures the *social sensitivity* of person i. Let's call the model in equation (2.1) the *baseline model*. In each round n, everybody who is still sitting considers her propensity $P_i^{(n)}$ and then decides, by a chance mechanism, whether or not to stand up.

Under the conditions of this model, the Borel-Cantelli Lemma implies that the number of standing people will converge to 1, as n goes to infinity.[2]

[2]This is easily demonstrated for any single agent i as follows: In each round k, $P_i^{(k)} \geq (1 - \sigma_i)\, q_i$. Thus, the probability that the agent will remain seated after n rounds is lesser

But in practice, only a finite number of rounds will be played, and if the σ_is are sufficiently small, it may well be the case that not everyone stands up. In order to better study this model, let us now turn to a numerical analysis of the model.

2.1.2 Numerical Analysis

Our model (and its extensions, which we will discuss below) suggests a variety of numerical studies. In order to best investigate these cases, we turn to instantiating the models as agent-based computer simulations. The simulations were written in Netlogo 4.0.4.

In the agent-based simulations, 1089 agents are seated in a 33×33 grid, all facing the same direction, in order to represent individuals seated in a theater. As is standard in agent-based models, time is broken up into discrete steps. In each step of the simulation, seated agents independently assess whether or not they should stand. Their decision procedure is simply an instantiation of the equation previously described. As was noted before, agents who are standing remain standing in perpetuity. Each simulation is run until either every agent is standing, or 1000 steps have passed. If each step represents one second of actual time, 1000 steps represents nearly 17 minutes, which we consider to be the extreme upper end of how long a standing ovation might last. For the purpose of analysis, all simulations were repeated 100 times.

Complete Information

In our first set of simulations, we examined agents who could see the entire audience. Their position in the theater had no effect on what information was available to them. As such, agents all worked from precisely the same information about what others in the audience were doing.[3]

As we have previously noted, the baseline model guarantees convergence on a standing ovation. So instead of discussing whether or not a standing ovation occurs, we study the *speed of convergence*. In particular, we are interested in determining how the parameters specified in the baseline model

than or equal to $(1 - (1 - \sigma_i)\, q_i)^n \to 0$. Thus, the agent will eventually stand up with probability one. Since the group is finite, the group will almost surely stand up as well.

[3] In the complete vision case, the expected time to convergence can be calculated even without recurring immediately to simulations, but modeling the system as a Markov chain with phase transitions.

affect convergence times.[4] To examine these effects, we must consider the agents' intrinsic propensity to stand up, and the agents' social sensitivity in turn. We will first examine the effects of the intrinsic propensity on ovation convergence.

In our studies of the intrinsic propensity we held social sensitivity σ_i fixed at various values (Figure 2.1) so we could examine how the increase of the intrinsic propensity by itself affected rates of norm convergence. In general we saw that, as the intrinsic propensity increases, the time of convergence decreases. This is, of course, exactly what we expect from the mathematics of the model (2.1).

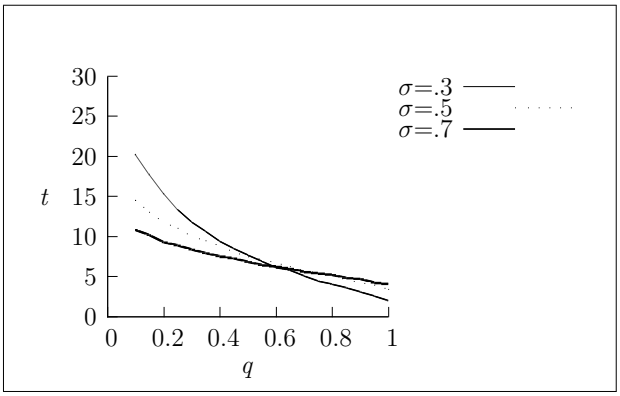

Figure 2.1: The effect of intrinsic preference on time to convergence

We found a similar story with an examination of the social sensitivity (figure 2.2). As before, what we saw is that as the social sensitivity increases the time of convergence decreases. However, the speed of convergence diminishes at a different rate when the social sensitivity is combined with a very low value of intrinsic propensity (q_i=0.1). This makes perfect sense: when both the intrinsic propensity and the social sensitivity are very low, it takes a long time until *everyone* is standing. It is enough that the social sensitivity slightly increases for the initial deadlock to be resolved.

We found that both parameters make a notable difference: each can cause the convergence rate to be significantly faster. However, the manipulation of

[4]For ease of analysis, we report on those models in which all agents have the same parameter values. We examined mixed populations, but did not find differences that merited separate presentation.

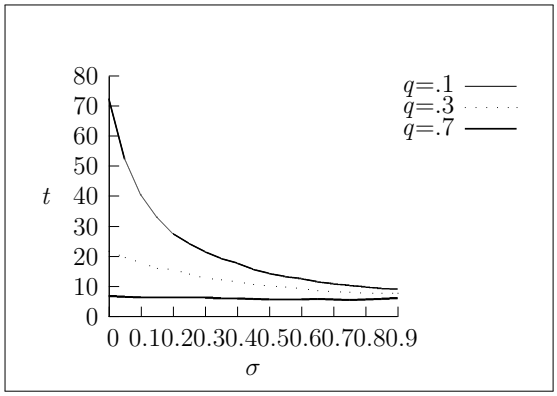

Figure 2.2: The effect of social sensitivity on time to convergence

social sensitivity appears to diminish convergence time more pronouncedly. This makes also sense: sensitivity to one's peers will accelerate any bandwagon effect as the population moves towards convergence.

Incomplete Information

In our second set of simulations, we examined agents who were limited in how much of the audience they could see. In particular, agents could only see the agents in front of them within their range of vision. In this model, agents could see all the way to the front of the theater, but only within a cone of 30 degrees. Thus, agents could not see anyone behind them, nor anyone outside of this limited scope of vision in front of them. This extension assigns different degrees of influence to the agents: those in the front rows are highly influential as their choices are taken into account by agents seated towards the rear, whereas the latter's choices affect few other agents. Notably, we obtain an asymmetry between information and influence: agents at the front can only see the behavior of a narrow peer group (or they don't care about the rest), while those at the rear have complete insight. This extension is therefore a particularly intuitive way to model mutual impact in social hierarchies.[5]

[5]The distinction between complete and incomplete vision is a variation in the environment of the model, whose effects have been explored in this paper throughout different specifications of the baseline model. Related results have been reported when significant. Variations in the environment of a model are thus orthogonal to variations in the structure of the models and to the study of their robustness.

To describe this more formally, consider an audience of R rows with L seats in each row. Now everybody takes only a fraction of the whole audience into account when calculating the extrinsic propensity, for example the cone of people in front of the person. Then, clearly, the behavior of the people in the first row will be more important than the behavior of those in the last row. After all, almost everybody will look at what the people in the first row are doing. In this case, the ratio $S^{(n-1)}/M$ is to be replaced by the expression

$$\frac{1}{|\mathcal{M}_i|} \sum_{j \in \mathcal{M}_i} s_j^{(n-1)} \tag{2.2}$$

Here $\mathcal{M}_i$ is the group of people person i can observe.[6]

As with our study of models of complete information, we first held the social sensitivity parameter fixed at discrete values to study the effects of the intrinsic propensity's increase (Figure 2.3). We found that the model with incomplete information behaved very similarly to the complete information case, though convergence times were notably slower at low values of q_i. Whereas in the complete information case, when $q_i = 0.1$ average time to convergence was 20.25, in the case of incomplete information average time to convergence was 59.47. As q_i grew, however, these disparities disappeared.

This suggests that while limited information can have a notable effect in slowing down convergence times in cases of low levels of intrinsic propensities, this effect rapidly diminishes as agents' intrinsic propensity to stand increases.

As we turn our attention to the effect of the social sensitivity parameter however, we find that limited vision has a strikingly large effect that reverses the trend seen in the model of complete information (Figure 2.4). Whereas before, moving σ_i from a low to a medium value induced significantly lower convergence times, in the model of incomplete information, as σ_i increases, convergence time also increases. Once $\sigma_i > 0.5$ we find particularly dramatic increases in both convergence times and variance. Limited vision of others has a dramatic effect on convergence, and for good reason. As the social sensitivity increases, the effect of the intrinsic propensity diminishes. When only very few agents are in a position to affect the behavior of others, it can easily happen that they remain seated. This can then lead those that look

[6]For modeling simplicity, we assume that each agent counts herself as a spectator. In this way, we avoid the issue that the model would otherwise be undefined for agents in the front row.

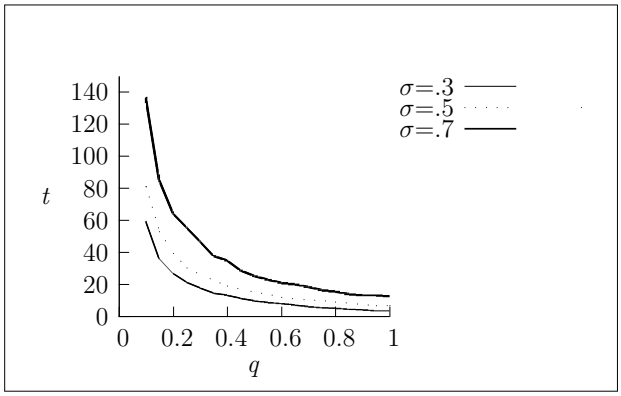

Figure 2.3: The effect of intrinsic propensity on time to convergence in the limited vision case.

to them for guidance about standing to also remain seated. This dynamic substantially dampens the bandwagon effect that is found in the baseline model with complete information.

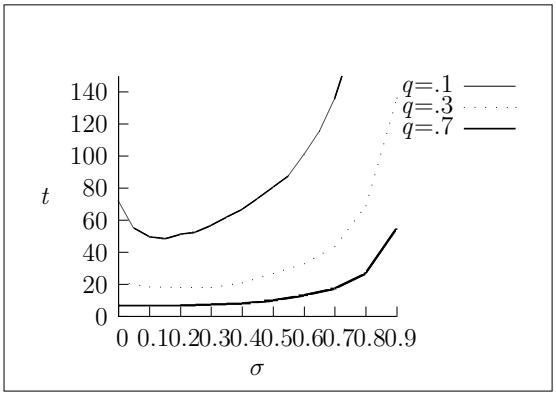

Figure 2.4: The effect of social sensitivity on time to convergence in the limited vision case.

2.2 Problems with the Baseline Model

While the baseline model helps to illuminate the basic structure of individual decisions that can result in the emergence of a descriptive norm such as a standing ovation, there are several reasons to suspect that the model is not yet adequate. Drawing upon philosophical and empirical literature on norm compliance (Bicchieri 2006; Young 2009), we can levy three major criticisms at the baseline model. We will look at each in turn.

The first way in which our model falls short is that it is not very sensitive to a more nuanced psychology of decision-making. One thing it fails to capture is the idea of entrenchment – people can often become set in their ways over time, and become less and less willing to change their minds, even if social influences become significant. Additionally, the baseline model lumps the notion of social sensitivity in with the notion of social contagion: it assumes that larger and smaller groups exert the same amount of social influence over a person's decision. But it is likely that in some instances, small groups are sufficient for influencing individual choices, while in others, a much larger group is necessary to change an individual's decision.

The second way in which our model falls short is that it assumes that the amount others can influence us is always constant across different contexts. But this is unlikely to be the case. In instances where one has strong preferences, it is likely that social pressure is less important. Whereas, when someone is fairly indifferent between two actions, social pressure might be the main determinant of that person's choice.

The third way our model falls short is that it makes significant structural assumptions about the nature of descriptive norms that may inhibit its ability to be a useful general model. This comes in two ways. Most obviously, the model always leads to a convergence on everyone standing. This is a highly suspect assumption: there are many concerts in which standing ovations fail to occur, just as there are many *potential* descriptive norms that never come to be. Even still, we can expect many situations in which *some*, but not all, agents take on a particular action, and for this to be stable. As it currently stands, our model cannot capture this fact. Additionally, the model suffers from having a built-in implicit assumption about the directionality of norms. In our baseline model, people go from sitting to standing. It is impossible for sitting to become a norm. Likewise, it should be possible for no norm to emerge.

In the following sections, we will present extensions to the baseline model that will in turn seek to address these three deficiencies. What we will show is that the qualitative results from our baseline model continue to hold as we investigate the first two deficiencies. As we explore the structural assumptions, we will find additional constraints on the emergence of descriptive norms that further enrich our account.

2.3 The Inertia Model

In this extension, we seek to address the lack of nuance in the psychology of the baseline model's decision procedure. To do this, we make two changes: First, there is a scaling factor $e^{-\alpha_i n}$, $0 < \alpha_i < 1$ such that the more rounds have passed, the less likely it is that someone stands up. This allows us to more carefully investigate the notion of entrenchment. Second, the propensity to stand up as a function of the others' behavior $S^{(n-1)}/M$ is taken into the β_i^{th} power, $\beta_i > 0$. All this can be represented by the following equation:

$$P_i^{(n)} = e^{-\alpha_i n}\left(\sigma_i \left(\frac{S^{(n-1)}}{M}\right)^{\beta_i} + (1-\sigma_i)\, q_i\right) \tag{2.3}$$

β_i can be thought of as a measure of *contagion* in the group: The smaller β_i ($0 < \beta_i < 1$), the higher the chance that a few isolated individuals who rise from their seats will affect the rest of the group. In other words, if we keep the number of agents following the norm fixed, the propensity of an individual to follow a norm is higher for smaller β_i. This reflects the fact that there are circumstances in which it takes a few agents to trigger a conformity effect than others. β_i determines the relative influence of the first agents adopting the norm vis-à-vis those agents that adopt it at a later stage. Thus, agents with a large β_i act on the basis of their propensity and the observed behavior of a crowd (as opposed to being responsive to the behavior of individuals and small groups). The break-even point is $\beta_i = 1$.

In this model, there is a nontrivial probability that not everyone stands up, even if infinitely many rounds are played.

Finally, it should be stressed once more that contagion and social sensitivity play different roles: while social sensitivity balances an agent's individual preferences against the impact of the behavior of others, the contagion parameter determines the rate at which the influence of an additional person

standing declines (or increases) with the number of people standing. If β_i is low the first few people standing will have a much larger influence than the final few.[7]

2.3.1 Numerical studies

The inertia model is meant to provide a mechanism for non-complete ovations, by providing a countervailing force on individual decision-making, encouraging some to remain seated. This is done with a time-dependent scaling function, which can be made more powerful by increasing the size of the inertia parameter α_i. This is done with a time-dependent scaling function, which can be made more powerful by increasing the size of the inertia parameter α. Secondly, it modulates the effect of social influence – by taking the social sensitivity component of the base model and raising it to the β^{th} power. We will investigate each of these modifications to the base model in turn, considering their effect on ovation size in equilibrium.

As the model description indicated, the inertia parameter α_i has by far the largest effect on ovation size. Here we will consider α_i with β_i fixed at 0.1. In general as the inertia parameter grows, we find an exponential decay in equilibrium ovation size. We examined values of $0.01 \leq \alpha_i \leq 0.5$ in steps of 0.01. As represented in figure 2.5, we find a rapid decay in equilibrium ovation size.

The inertia parameter controls the rate at which agents are willing to stand as time goes on, which heavily dampens their ability to respond to new information as it is revealed to them. As agents become increasingly stubborn as time goes on, this limits their interest in standing regardless of what anyone else is doing. We find a similar story for increasing values of the contagion parameter β – since individuals respond less and less to smaller groups for higher β – it is more difficult to get a bandwagon effect initiated even if they are increasingly sensitive to large groups. The large groups simply cannot form if smaller groups do not have sufficient attractive force. In this way, α_i and β_i work in concert to limit ovation size: β_i limits the power of an initial small group standing, and then α_i increases the stubbornness

[7]We have also explored a different extension of the model with a counter-force to the overall conformity. This second way assumes that some people increasingly resist standing up as more people stand. They act against the mainstream. We do not present this non-conformist model, as we did not find a significant deviation from the baseline model, even if this condition may be psychologically relevant.

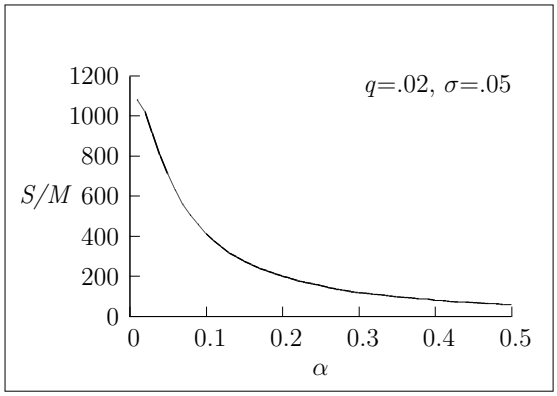

Figure 2.5: The effect of the inertia parameter on the number of people standing

of agents sitting as the groups get slightly larger over time. This combined effect can neuter a group's ability to create a social bandwagon.

In the inertia model, we find that the imposition of incomplete information has very little effect on how the active variables in the model affect ovation size. In the case of α_i, we find no discernible difference between the complete information model and the model of incomplete information. In the case of β_i, we find few differences where $0 < \beta_i < 1$.

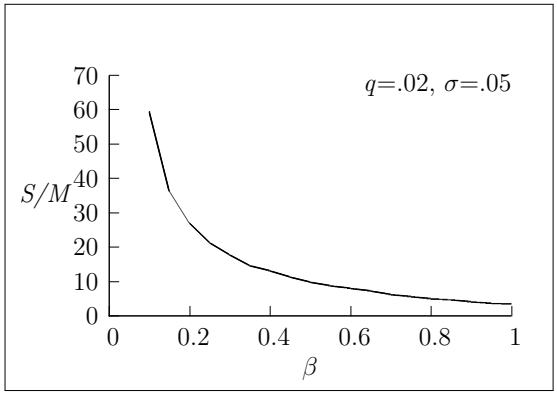

Figure 2.6: The effect of a contagion parameter on the number of people standing

2.4 The Endogenous Social Sensitivity Model

The baseline model and its initial extension, the inertia model, consider social sensitivity as an exogenous parameter: It is a parameter that balances one's intrinsic propensity to comply with the behavioral rule with the impact of group behavior. For an agent with high social sensitivity, the impact of group behavior will dominate the impact of one's individual judgment on the quality of the concert, and vice versa, for an agent with low social sensitivity, group behavior will have a minor impact on the agent's decision. Social sensitivity does not, on that account, depend on one's intrinsic propensity or the number of people already following the behavioral rule.

This view can, however, be challenged. In her book *The Grammar of Society*, Bicchieri (2006) has shown that *empirical expectations* of the behavior of others are crucial to whether descriptive norms emerge and persist. If an agent expects a critical part of the population to follow a behavioral rule, then she will most likely follow the rule as well. Further, an agent may only become aware of the existence of a candidate alternative norm once it is sufficiently widespread in the population. If a large part of the group starts to comply with the rule, the agent reasonably expects that the behavior will spread to the entire group, and these expectations overrule an agent's independent preferences as a determinant of her individual behavior. Conversely, if the percentage of individuals abiding by the rule is lower than such a critical value, group behavior barely affects individual behavior. Social sensitivity should thus be treated as an *endogenous* variable crucially depending on the observed behavior in the group.

Both in the baseline case and the inertia extension, social sensitivity was considered exogenous. If it is low at the outset, then it will stay low, even if the norm spreads rapidly in the group. This delays the convergence process. It is therefore worthwhile to investigate whether our results remain robust under the feedback effects described above. To that end, we have to specify the dependency between the variables of the original model.

We keep the baseline equation (2.1) intact and only write social sensitivity σ as a function of the other variables.[8] As argued above, social sensitivity should be very low ($\approx \varepsilon$) if S/M is significantly below a critical value, and very high (≈ 1) if S/M significantly exceeds that value. It is natural to

[8]This implies that σ is time- and agent-dependent, but for reasons of simplicity, we drop the subscripts in this exposition.

assume that the lower the intrinsic propensity q, the higher the threshold: If an agent strongly dislikes the behavioral rule, her empirical expectations of compliance with the rule will be higher, and the group will have to behave more homogenously in order to meet them. Thus, we might choose

$$\sigma = \begin{cases} 1, & \text{if S/M} \geq 1 - q, \wedge P = q \\ \varepsilon, & \text{otherwise.} \end{cases} \tag{2.4}$$

Figure 2.7 below gives a graphic representation of the model. On the x axis is the number of people standing up in the total audience, on the y axis the values of sigma.

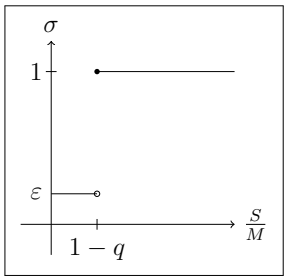

Figure 2.7: Discontinuous Model

When the intrinsic propensity of an individual is high her reliance upon others for the decision to stand up or not is triggered by few individuals standing, whereas when it is low she needs to see more people to follow them as well.[9]

2.4.1 Numerical Studies

In this set of simulations, we studied the effect of the intrinsic propensity on time of convergence. We examined values of q varying from 0.01 to 1 in steps of 0.01 For simplicity of our treatment, we do not introduce the inertia and contagion parameters, as we have previously examined them in isolation.

[9]It is, however, not clear whether real social sensitivity is as discontinuous as this equation suggests. It seems more realistic that in many cases people have moderate individual preference coupled with moderate social sensitivity. So we 'smooth' the function by introducing an additional parameter that governs the quickness of the transition. We did not find that this variation had a significant difference on the final result, so we only report on the discontinuous case, as the mathematics are more straightforward.

Further, as we didn't find crucial differences between the incomplete and the complete information case, we only present the first case. As expected, and the figure below shows, we find that the time of convergence decreases for increasing values of intrinsic propensity.

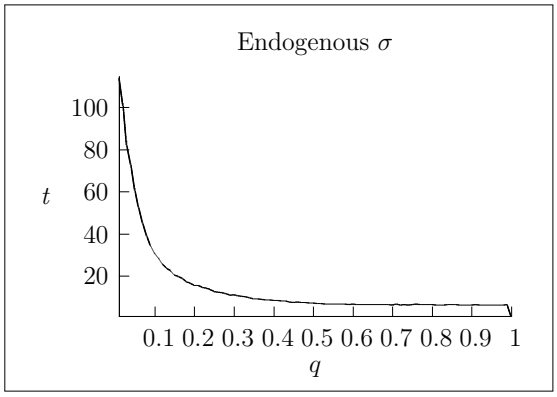

Figure 2.8: The effect of the endogenous sigma on the time to convergence

2.5 The Symmetric Model

In our final model, we consider a generalization of our original model. As we have previously discussed, the models we have been considering thus far all have an implicit directionality: people start out sitting, and potentially stand. The only descriptive norm that can emerge is one of ovation. But this assumption limits our ability to describe the emergence of descriptive norms more generally. As an example, consider the norm that governs how forks are used while eating. In Europe, a fork is used in the left hand, so as to enable the eater to use a knife in her right hand. In the United States, however, while forks are held in one's left hand while cutting food, they are then moved to the right hand for raising food to one's mouth. While either method of using forks is perfectly suitable for consuming food in a polite and efficient manner, they are regionally segregated. In the United States, it is rare to see the European method, and the US method is rarely seen in Europe. What we can notice is that there is no particular reason to think that one method is prior to the other, so we cannot model this as a standing ovation. So

how might we provide an explanatory framework for the emergence of norms when the potential behaviors are on equal footing?[10]

What we propose is a return to our baseline model (2.1), but with a few crucial changes. First, we introduce two agent types, each type having a preference for one of the two actions available to them. So type 1 agents prefer action 1, and type 2 agents prefer action 2. Second, we re-interpret the variable $0 < q < 1$, such that 0.5 represents the indifference point, rather than 0. On this new interpretation, 1 represents a strong preference for the action of one's type, and 0 represents a strong preference away from this action. Third, we allow agents to change their minds: whereas in previous models once an agent has chosen to stand, she must remain standing, in this model agents can reverse course and go back to their previous action. Finally, when we initialize the model, agents are randomly (and independently) assigned an initial starting action. So, unlike previous models, our starting state has half of the agents performing the first action (say, the European way of using forks), and half the agents performing the second action (like the US method of using forks). These changes allow us to investigate several things that could not be examined in previous models. In particular, since we are treating the two methods of using forks as symmetrical to each other, and we allow agents to change their minds, we should expect the models to exhibit more complex behavior. More importantly, however, we are now able to clearly separate cases of norm emergence from behavioral regularities, since we have differing preferences amongst agents.[11]

The model's equilibrium states can be broken down into three classes: descriptive norm emergence, large-scale behavioral regularities, and a mix of behaviors. Descriptive norm emergence is found when the entire population settles on a single action. In these cases, half of the population must be going against their intrinsic preferences, and instead their social sensitivity drives their decision-making. This can be contrasted against large-scale behavioral regularities, which are cases in which all agents of one type choose the same

[10]Young (2009) develops a model of innovation diffusion that assumes priority of one action over another, that shares some characteristics with Schelling (1971). While this model is rather elegant, it does not capture the possibility of equal footing for either norm, or the possibility of no norm emerging.

[11]Recall that while descriptive norms rely on agents being motivated out of a desire to do what others do, behavioral regularities are simply cases in which individuals all perform the same action, but for independent reasons. They just all happen to prefer the same action.

action, but choose a different action from agents of another type. So everyone who prefers European fork-handling employs it, and likewise everyone who prefers US fork-handling does so. In this case, we claim that individual preferences are the most powerful guide to decision-making, and so social sensitivity effectively drops out of consideration. Our final case is what is left over: a mix of influences, none of which is strong enough to completely guide behavior. In these cases, both preferences and social sensitivity are at work, neither of which is sufficiently strong to overpower the other. So we find an unsystematic mix of behaviors.

Let us now turn to a numerical analysis of this model.

2.5.1 Numerical Studies

This model requires a different approach to our analysis. Rather than consider something like time to convergence, we must instead consider the probabilities of settling into the three different states for the different values of intrinsic preference and social sensitivity. The initial state is shown in the figure below, as it appears in the Netlogo interface. The grid represents the theatre audience, composed by two agent types: squares are agents who prefer standing and circles are agents who prefer sitting. White squares stay for those agents who perform their preferred action, in this case standing, otherwise they are black. Black circles stay for those agents who perform their preferred action, namely sitting, otherwise they are white.

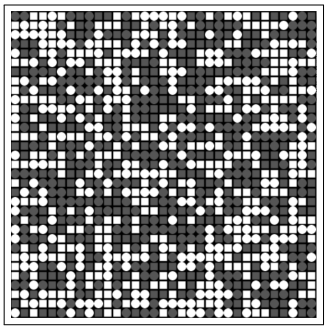

Figure 2.9: Symmetric Model's Initial State

In the set of simulations for the symmetric model we examined the probability for a norm to emerge according the distributions of social sensitivity

and intrinsic propensity between the audience. The norm emergence corresponds to a state in which all agents in the audience perform the same action, regardless of their intrinsic propensity. Graphically (figure 2.10), this happens when all circles and squares are white (or when all circles and squares are black).

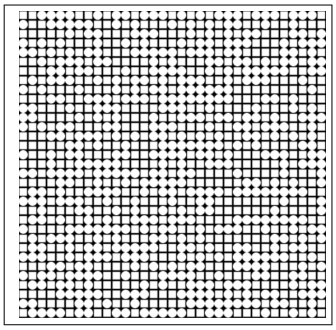

Figure 2.10: Full Norm

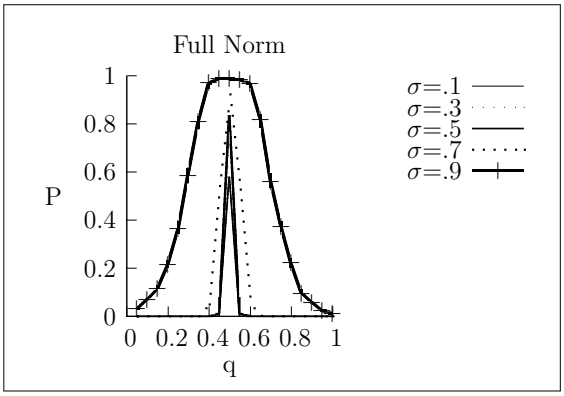

Figure 2.11: The probability of descriptive norms' emergence

What we found was that the emergence of full descriptive norms is quite rare. Figure 2.11 represents the probability of descriptive norms emergence for increasing values of intrinsic propensity. Each curve corresponds to fixed values of social sensitivity. We can see that this probability increases as the agents become indifferent between the two actions ($q \simeq 0.5$) and in general for higher values of social sensitivity.

More common are large-scale behavioral regularities. These occur when agents perform their favorite action, e.g. when all squares are standing and all circles are sitting. These outcomes can happen for wider ranges of social sensitivity, so long as the intrinsic preference is more extreme in value. Figure 2.13 shows that the probability of behavioral regularities is lower for intermediate values of intrinsic propensity and for lower values of social sensitivity.

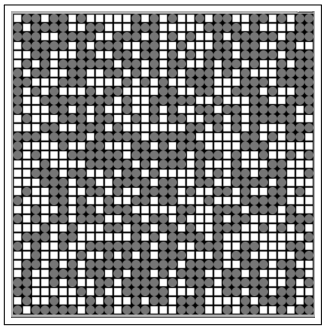

Figure 2.12: Behavioral Regularities

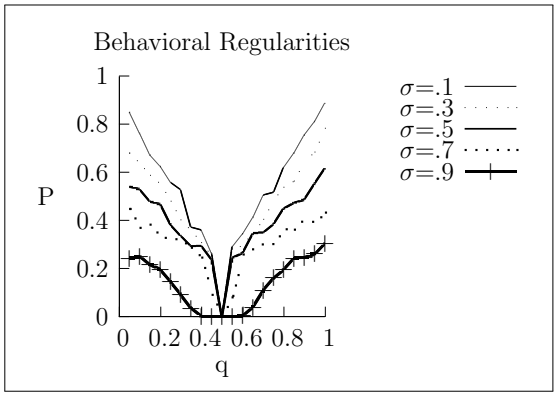

Figure 2.13: The probability of behavioral regularities' emergence

Most common of all, however, are mixed outcomes, those in which some of the agents perform their preferred outcome and others don't (graphically, this corresponds to a situation similar to the initial state but with a different

distribution of colors, according to those circles and squares that modified their initial state).

This result seems to comport well with the real world: though descriptive norms are quite common and are found in a very wide variety of social situations, there are many more possible descriptive norms than there are actual descriptive norms. Most of our day-to-day behaviors are not norm-governed, even though many are.

2.6 Conclusions

We have argued that a model of standing ovations can provide a useful framework for the investigation of the emergence of descriptive norms. While we do not claim that all descriptive norms have the character of standing ovations, we have tried to suggest that with a few modifications, we can transform a model of standing ovations into a general model of the emergence of descriptive norms. In particular, we wish to stress the qualitative match of results across the various models we present. The baseline, inertia, and endogenous social sensitivity models all explore the convergence dynamics of a 'directed' transition from one behavior to another. Though they build in substantively different psychological assumptions about the agents involved, we find that these large perturbations do not shift us far away from our original baseline results.

While descriptive norms themselves most often are not fully captured by the baseline model, it can often be the case that these sorts of directed transitions *can* describe the propagation of information about the social context for behaviors. For example, Christians remove hats in church to show proper deference, but not at sporting events. When they enter a novel environment, for which they may or may not have to signal deference – say, going into a classroom or a museum for the first time – they may look to others for signals of what they should do. When we enter into a novel situation, we may not be sure which of our already-established norms ought to govern our behavior. A directed transition model, like a standing ovation, might be a good representation of this sort of phenomenon.

The final model we consider, we contend, does capture the essential elements of the emergence of descriptive norms, given that it is possible for any behavioral rule, or none at all, to emerge as a norm. What is so striking about this last model is that it is only a minor modification of the original

baseline model, but provides a dynamic that displays the relevant considerations for the potential emergence of a descriptive norm. We did not do anything to change the fundamental decision procedure – we simply allowed people to change their minds, and have preferences for more than one action. But with these small changes, we generalized the model, and enabled ourselves to discuss a much larger class of social phenomena.

This kind of modeling offers some insights into the nature of descriptive norms that might be difficult to arrive at otherwise. In particular, what we find, especially by studying our symmetric model, is that whether a norm emerges at all, let alone which norm it is, is remarkably contingent on factors that have nothing to do with the substance of the norm itself. Whether it is table manners or audience behaviors, or even how we dress, we do not follow them because they are somehow superior to their alternatives, but rather we follow them because a mix of social and personal factors happened to nudge us in one direction rather than the other. We often place value on these norms, but we should avoid making the mistake of thinking that this value comes from the action itself. Rather, we can see the value of an action coming from the fact that we have become accustomed to doing it.

In the next chapter, I will consider a different decisional rule for norm compliance, based on a Bayesian updating mechanism for belief revision. I will investigate the question of whether it is possible to provide a rational reconstruction of the individual behavior to conform to a descriptive norm. In so doing, an alternative explanatory framework will be presented, that grounds the inference to the existence of a norm in the same reasoning process we use to infer regularities in the natural world.

Chapter 3

Why are descriptive norms there?

3.1 Introduction

The previous chapter deals with various heuristic models for the emergence of descriptive norms.[1] But this leaves a challenge: why should we expect those heuristics? Is there some deeper justification that we can find for the social dynamics under investigation? As we have seen, descriptive norms are a curious class of behaviors: unlike social norms, there is no strong normative component. Unlike conventions, they aren't solutions to two-sided coordination problems.[2] Descriptive norms exist where individuals follow a common pattern of behavior simply because they have a preference for that behavior, if they think enough of the rest of the population follows it as well.[3]

By and large, descriptive norms don't need to exist at all – they aren't solving problems that social groups need solutions to. Even so, our lives are full of descriptive norms. Fashions, fads and all manner of trends fall under the category of descriptive norms. Opinion dynamics, and some ways of

[1] This chapter is based on Muldoon, Lisciandra and Hartmann (under review).

[2] Descriptive norms can be understood either as one-sided coordination problems, or as creating them in a similar manner to the (Bicchieri 2006) account of social norms transforming mixed motive games into coordination games. Unlike a convention, which provides a solution to a two-sided coordination game, there is no need to coordinate expectations across parties. One party can simply choose to match what others do.

[3] We follow Bicchieri's definition (Bicchieri 2006). More formally, a descriptive norm is a behavioral rule R for a population P in a context C where individuals in P have a conditional preference to follow R if they believe that a large enough proportion of the population P follows R in C. This belief is their empirical expectation of rule compliance.

expressing political or religious support, can follow the patterns of descriptive norms. These norms emerge from our social interactions, continuously and spontaneously.[4]

Precisely because descriptive norms do not represent solutions to problems that social groups face, a distinguishing feature of descriptive norms is they are essentially unconstrained in content. There isn't an underlying fact about the behavioral rules themselves that will guide a population to one descriptive norm versus another, because either is equally good. This suggests that the reasons a particular descriptive norm is 'chosen' is independent of its substance. This is rather different from other previously studied social dynamics, such as the diffusion of a new technology (Young 2009), where features of the choices help guide the pattern of adoption. A new fashion trend isn't typically adopted because of pragmatic features of the new fashion. For instance, 'Brown is the new Black' is a rather different claim than, say, pointing out that word processors make writing and editing documents easier. While some people may prefer brown due to personal tastes prior to any new trend, that is not quite the same as recognizing the innate superiority of brownness over blackness, in the way that we may want to argue that word processors are just superior to typewriters. Quirks of personal taste may leave some in favor of typewriters for a time, but word processors end up taking over because of their efficiency and additional capabilities. In the case of color choice between brown and black, there is no such outside objective measure that can push people to one color or the other. In the case of descriptive norms more generally, we can understand the process of change as being governed purely by the process itself. The chosen norms need not offer any advantage over their alternatives.

In what follows, we propose that the emergence of descriptive norms, and their apparent arbitrariness and instability, can in fact be understood in light of a larger epistemic apparatus. Within this framework, descriptive norms are seen as a byproduct of a domain-general mechanism of hypothesis

[4]The distinction between descriptive norms and other informal norms, such as social norms or conventions, however, has fuzzy boundaries. Even in trivial cases, like following a fashion trend, a few individuals might follow them out of fear of punishment or of social rejection, even if no one else would even think of punishing them. Individuals can have a variety of motivations – sometimes one subset of people have normative expectations (the belief that others think you ought to follow the rule), while the rest of the population only has empirical expectations. For our purposes, we focus only on canonical cases of descriptive norms where people lack second-order normative beliefs.

generation and belief revision. In particular, we argue that our disposition to find and follow rules stems not from their immediate utility, but rather from the immense value that the general epistemic apparatus has in our lives. That is, while the general disposition to discover rules and act on the basis of our knowledge of them can be utility enhancing, this does not mean that each instance of rule-following must itself be justified on such grounds. Any given descriptive norm may well be arbitrary, even if the general process that creates them is not.

To describe this process, we imagine that an individual finds himself in a situation similar to that of a scientist who is looking for the evidence in support of his hypothesis. As the rational way to proceed in order to estimate the probability of a certain hypothesis about the world is by Bayesian updating, similarly, we express the individual decision problem as a conditional probability. Accordingly, the individual's degree of beliefs in an action being a norm is a function of the evidence of other people's behavior and their reliability. In other words, unlike in the previous chapter, here agents are Bayesian updaters – they have a domain-general reasoning system that helps them to update their hypotheses about how the (social) world around them operates in light of new evidence.

Notice that the specific interest in descriptive norms is twofold: first, these are norms which involve only one level of expectations – namely, what an individual believes the majority of people will do in similar situations. This allows for a formal model of their emergence to remain easily treatable; secondly, the philosophical question about this type of norms is whether – at least under some conditions – rational agents will comply with them, given the behavior of the members of their group, even if there is no objective reason to do so. When dealing with descriptive norms, the hypothesis under consideration, namely whether the norm is worth following or not, is probably neither right or wrong *per se*. In cases like these, the evidence provided by other people's behavior becomes the only element upon which to rely. There is no external or objective source of information about the hypothesis, beyond the actions of others.[5] Within an epistemic framework we can analyze the way in which this evidence is processed by rational agents given a domain-general updating system, and we can see the consequences of that process.

[5]Recall the 'Brown is the new Black' claim versus 'word processors are superior to typewriters' – we can have efficiency measures to compare the machines, but we would be at a loss for an equivalent measure for the colors.

This chapter will proceed as follows: First, we present an analytic model of an individual's reasoning that we employ. Next, we will describe a computer simulation that implements this for a group of agents that have some structure to what they can discover socially. We then analyze the results of the simulation, and argue that this model provides evidence that we can make sense of the emergence of descriptive norms if we see it as an instance of a larger cognitive apparatus that helps us be responsive to evidence. Finally, we argue that this practice of more domain-general idealized reasoning – using very strong rationality assumptions that likely go beyond our cognitive capacities – allows us to see more universal dynamics across a number of social situations.

3.2 The Model

When studying the emergence of social behaviors, we need a formalism that can account not just for what happens at a group level, but how individual decisions lead to a group outcome. In this study we rely on a Bayesian approach, primarily for its ability to carefully monitor what happens at the individual level.[6] With a Bayesian model, we can express the individual's degrees of belief about a certain hypothesis and elaborate it with the laws of probability calculus. Broadly speaking, Bayesian epistemology deals with the logic of inductive reasoning and expresses formally how we should learn from experience and how we should estimate our hypotheses under conditions of uncertainty (see Hartmann and Sprenger (2010) for an introduction to Bayesian Epistemology).

The mathematics of Bayes' rule is straightforward. Its main components are *the priors*, namely the probability of the hypothesis before the evidence, and *the likelihoods*, which measure the probability that the evidence supports the hypothesis. By Bayes' rule, these components are used to compute the *posterior probability*, namely the probability of the hypothesis conditional on the empirical evidence.

As a domain-general updating process, Bayes' rule can be applied to a variety of situations. According to the context, the priors and the likelihoods express the role of different pieces of information for our hypothesis. It

[6]Nothing in our argument relies on Bayesian updating in particular – we employ Bayesian updating because it is a well-understood, straightforward model of domain-general belief revision.

can deal with the probability of getting a red ball when drawing from urns filled with balls of different colors, or how much a positive test result should change one's beliefs about whether one has contracted a disease, or in legal settings, how conclusive DNA evidence might be in determining the guilt of a defendant. Across these cases, Bayes' rule shows us how much we can learn from new information and update our previous beliefs to arrive at a new belief about the probability of some hypothesis.[7]

In the same way in which Bayes' rule is used to reconstruct an ideal process of updating one's beliefs about a certain hypothesis about the physical world, in this study we adopt it to model changes in beliefs about hypotheses about our social world. The way we draw social inferences resembles the more general process of learning from experience. However, in the social world, evidence is interactive: we learn from each other's choices[8]. To model this situation, we suppose that the members of a group assign a probability to whether a given behavior is a descriptive norm on the basis of the priors, namely the probability of the hypothesis before the observed behavior and of the empirical evidence at their disposal. In this estimate, however, not all evidence is treated equally. We assume that not only are some individuals more reliable than others (in terms of reliably following norms when present), but also that different individuals are more or less sensitive to other people's behavior.

In what follows, we will formulate a Bayesian model of norm discovery. Our model aims to be a simplified representation of a social situation in which there are multiple behavioral patterns, at most one of which emerges over time as a descriptive norm. To do this, we have eliminated as many superfluous features of real social systems as we could to focus on the core dynamics. The model consists of n agents, representing some pre-existing social group. Within this pre-existing social group are two common behaviors: one which we label $\mathcal{N}$ – the behavioral rule we consider as a possible descriptive norm – and its alternative. The model is agnostic about the content of these behaviors. We choose one as $\mathcal{N}$ without any loss of generality. The model treats both behavioral patterns symmetrically. The model unfolds over time. We divide time into discrete rounds, and all agents complete the decision process exactly once per round. In what follows, we will formally describe that decision procedure and how it drives the model.

[7] See Hacking (2001) for a formal treatment of the aforementioned examples.

[8] See Schelling (1978) for typical interactive, critical-mass models in the social sciences

Each agent has beliefs about the state of the world with respect to the proposed behavioral rule. We represent this formally as follows. Each agent has a propositional variable H, which can have two values. 'H', means that the behavior is a descriptive norm. '¬H' simply means that it is not a norm. Furthermore, agents can have beliefs not just about the status of the behavioral rule as a descriptive norm itself, but about what others are doing. The variables $E_i^{(k)}$ (with $i = 1, \ldots, n$) have two values: $\mathrm{E}_i^{(k)}$: Group member i follows the proposed behavioral rule in round k, and $\neg\mathrm{E}_i^{(k)}$: Group member i does not follow the proposed behavioral rule in round k.

In round 0, everybody is equipped with a probability function $P_i^{(0)}$ and starts with a prior probability of H, i.e.

$$P_i^{(0)}(\mathrm{H}) = q\,, \tag{3.1}$$

with $i = 1, \ldots, n$. We call q the *intrinsic propensity* to follow the norm. This is simply the agent's internal preference for the behavior, independent of what anyone else does. For reasons of simplicity, we assume that all group members have the same intrinsic propensity. This assumption can easily be relaxed. On the basis of the prior probability and the epistemic sensitivity (more on this parameter below), group member i will follow $\mathcal{N}$ or not.

At the end of round 0, each agent has a *profile* $F^{(0)}$ about other agents' decisions $:= < \mathrm{E}_1^{(0)}, \neg\mathrm{E}_2^{(0)}, \ldots, \mathrm{E}_\mathrm{n}^{(0)} >$. We define the *profile relevant for group member* i as $F_i^{(0)} :=< E_k^{(0)} : k \in \mathcal{C}_i >$. Here $\mathcal{C}_i$ is the set of labels of the group members that group member i takes into account (e.g. the labels of those group members in the visual range of i).

In round 1, each group member updates on the profile relevant to herself, i.e.

$$P_i^{(1)}(\mathrm{H}) = P_i^{(0)}(\mathrm{H}|F_i^{(0)}). \tag{3.2}$$

By Bayes' Theorem (see Bovens and Hartmann (2003) ch. 3), we obtain

$$P_i^{(1)}(\mathrm{H}) = \frac{q}{q + (1-q)\cdot l_i^{(0)}} \tag{3.3}$$

with the likelihood ratio $l_i^{(0)}$

$$l_i^{(0)} = \frac{P_i^{(0)}(F_i^{(0)}|\neg\mathrm{H})}{P_i^{(0)}(F_i^{(0)}|\mathrm{H})}\,. \tag{3.4}$$

Let us now calculate the likelihood ratio. To do so, we first make the following independence assumption:

$$E_i^{(k)} \perp\!\!\!\perp E_j^{(k)}|H \tag{3.5}$$

for all $i, j = 0, \ldots, n$ and all rounds $k \geq 0$. The assumption is as follows: If we know that the behavioral rule is a descriptive norm (or not), then we will learn nothing new about the question as to whether group member i follows the norm if we learn that group member j follows the norm. What group member i does depends only on the truth value of H. A further justification for the conditional independence is that people might simply assume that the individuals act on the basis of the norm and not on the basis of what others are doing. In other words, within a subjective Bayesian framework, it is sufficient that individuals assume conditional independence for the assumption to apply.[9]

Next, we assume that

$$\begin{aligned} P_i^{(k)}(\mathrm{E_j^{(k)}}|\mathrm{H}) &= c \\ P_i^{(k)}(\neg\mathrm{E_j^{(k)}}|\neg\mathrm{H}) &= c\,. \end{aligned}$$

Here c is a parameter that measures the *expected compliance*. That is, it measures to what extent group member i believes that group member j will follow the behavioral rule if it is a descriptive norm, or not follow the behavioral rule if it is not a descriptive norm.

With these assumptions and the definition

$$T_i^{(k)} = \sum_{l \in \mathcal{C}_i} E_l^{(k)} \tag{3.6}$$

(k refers again to the round in question) we can now calculate the likelihood ratio:

$$l_i^{(0)} = \left(\frac{1-c}{c}\right)^{2T_i^{(0)} - n_i} \tag{3.7}$$

where $n_i = |\mathcal{C}_i|$, i.e. the number of group members in the cone of group member i.

[9]It might be asked why an individual should assume conditional independence, given that that assumption is false for her. The main rationale is that, when deciding whether or not to follow a descriptive norm, individuals have the tendency to disregard their own responsiveness to other people's behavior. In other words, individuals tend to believe that they are among the first ones to adopt a certain new behavior, that they have not been influenced by other people, and that it is their personal taste for the object or the action in question that determines their choices. This is the sense according to which the independence assumption holds as a psychological motivation underlying the individual decisions to comply with descriptive norms.

These equations generalize to later rounds. In round $k+1$, all group members update according to

$$P_i^{(k+1)}(\mathrm{H}) = P_i^{(k)}(\mathrm{H}|F_i^{(k)}) \tag{3.8}$$

with the likelihood ratio

$$l_i^{(k)} = \left(\frac{1-c}{c}\right)^{2T_i^{(k)}-n_i} \tag{3.9}$$

In each round k, we assume that group member i decides to follow the behavioral rule if $P_i^{(k)}(\mathrm{H}) > 1-s$. Here s is the agent's *epistemic sensitivity*.

We assume that the group members continue to update their beliefs even if they are already following the behavioral rule. It is therefore possible that someone who followed the behavioral rule in round k will stop following the behavioral rule in round $k+1$. This is simply to better match the real world – fads can fade away over time. Once a behavioral rule has become a descriptive norm and has full compliance by a given social group, it may be difficult to move away from it, but it does happen with some regularity.[10] To better capture the possibility of these dynamics, we do not artificially limit the updating procedure to simply favor norm adoption.

The reader should notice that in this model, each agent applies the standard Bayesian belief revision machinery to the particular case of norm adoption. We suppose that this belief revision machinery is around for other aspects of one's life – it is present in our social reasoning because it is used generally when we reason about the world. We enhance this general model by supposing that there are a few specifically social characteristics of our reasoning that must also be taken into account. These will be discussed in more detail in the next section. However, we note that the small addition of these parameters is all that is necessary to transform the basic reasoning process of Bayesian updating into a social rule discovery mechanism.

[10] Grunge clothing was popular for several years before it largely disappeared. Bangs are sometimes widely adopted, and then disappear for a while. Text messaging has largely supplanted once-dominant phone calls for quick messages amongst friends.

3.3 Simulating the Model

In order to best explore the model's dynamics, particularly with larger social groups, it was necessary to implement the model in an agent-based simulation. Agent-based simulation allowed us to use a relatively large population (1089 agents), and investigate somewhat more realistic representations of peer influence. For the purposes of this study, we imagine agents to be sitting in a 33×33 grid, with everyone facing forward.[11] Each agent sees the agents in its front visual cone (See Figure 2.1). The intuition behind this representation is that each agent's information and influence is position-dependent. The farther back one is, the more information they have, since they can see more of the other agents. However, the closer one is to the front, the more influence one has, as they are more likely to be seen by others. This structure allows us to represent real-world hierarchical relations in social groups in a general way. This is motivated by work using models of standing ovations as representations of social influence (Miller and Page 2004).[12] For the purpose of analysis, all simulations were repeated 100 times. [13]

Each agent starts out by following either the proposed behavioral rule or its alternative. This choice, since it relates to a particular action, is fully visible to others.[14] Because of this, agents can reliably update their beliefs.

[11]We implemented this simulation in Netlogo 4.0.4. The grid size is the simulation software's default setting. We explored smaller grids, and did not see qualitative differences. We report on this population size as a compromise between the desire for a large social group, and the super-linear increase in computational costs (in terms of time) of the simulation as more agents are added.

[12]While many other network structures are possible, we focus on this approach. There is not a significant qualitative change if we use a more standard Von Neumann or Moore 8 neighborhood. What primarily drives the results is the agents' limited information. The assumption of local information and some social hierarchy in how descriptive norms emerge is based on the consideration that full information models are extremely unrealistic – very rarely in our social lives to we have complete social information about an entire extended social group.

[13]The three main parameters, i.e. intrinsic propensity, expected compliance and epistemic sensitivity are bounded between 0 and 1 and in the simulations we observed the effect of the variation of one parameter, while keeping the two other fixed on medium, low, or high values.

[14]Think of clothing fashions, for example. Descriptive norms, especially ones that have any longevity, have to be associated with some public display or action, otherwise empirical expectations cannot be coordinated. Since there is no normative aspect, there is no reason to a have a descriptive norm about private behavior. Outside of actions influenced by our normative expectations, private behaviors do not have social motivations.

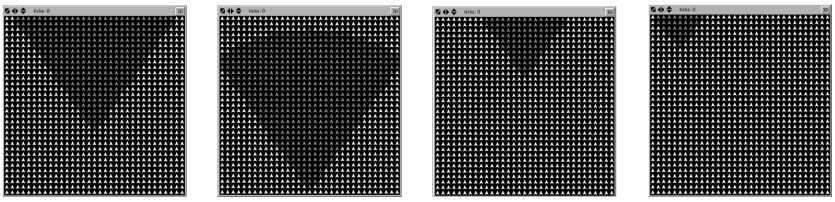

Figure 3.1: Series of grids representing different agents' visual fields. The grey cones in each grid show the visual field of the agent at the vertex of the cone.

We investigate what happens as agents update their beliefs over time. Under what conditions should we expect norm emergence? Our investigation revolves around the three free parameters of the model. These parameters augment the standard Bayesian approach by injecting social aspects of our reasoning. These social aspects are the *intrinsic propensity* of an agent to follow the proposed behavioral rule, an agent's assessment of other agents' *expected compliance* to the behavioral rule, and each agent's *epistemic sensitivity*.

More specifically, the **intrinsic propensity** measures the individual preference to follow the behavioral rule regardless of other people's behavior. This parameter (combined with the epistemic sensitivity) affects the initial conditions of the model. We use these parameters to determine the initial distribution of agent behaviors. In subsequent rounds, agents update on the evidence provided by other individuals according to the Bayesian procedure described above. If we reflect on the meaning of these parameters, we ought to expect that by and large the average individual would be fairly neutral between choices of action, since there is no particular utility benefit to either action. In the case of descriptive norms, people's priors should not be that strong in either direction. As such, we should expect that instantiations of the intrinsic propensity parameter should be somewhere in the middle of the range from 'absolutely in favor of the behavioral rule' and 'absolutely against the behavioral rule'. In such cases of relative indifference, the other two parameters of the model matter more to an individual's choice: expected compliance and epistemic sensitivity. However, there are scenarios in which we might expect strong individual preferences for given behavioral rules. Due for example to their past experience, individuals might behave according to consolidated practices, that they bring along once in a new group. If the pro-

posed behavioral rule happens to link to an individual's larger set of views or preferences, we might find more extremal values for intrinsic propensities. For instance, some people like wearing plaid shirts, regardless of whether there is a larger grunge fashion trend. Some might think that text messaging is a distasteful form of communication, even if many others use it.

The **expected compliance** parameter measures the reliability that each agent assigns to the members of the group. It indicates if the source of evidence matters for the decision as to whether to follow the norm. In real-world scenarios, this corresponds to reliable or influential people who, for whatever reason, are considered to be competent on that matter. Hence, agents who have been assigned high expected compliance will be judged as reliable indicators of the presence of a norm if they are following it and of the behavioral rule's failure of becoming a norm if they are not following it. This helps capture the idea that we have potentially different assessments of the same evidence (following the behavioral rule) when it comes from different sources (more or less reliable trendsetters).

The **epistemic sensitivity** parameter measures an agent's individual degree of responsiveness to perceived empirical regularities. In other words, the epistemic sensitivity parameter is the means by which agents convert their epistemic state into a motivation for action. For instance, in a non-social case, epistemic sensitivity determines whether an agent would act on her belief that a particular river floods with some regularity. This may lead her to build her house farther away from the river's banks. In the social case, it reflects the fact that some agents are more responsive to social cues than others. Some people seek to match their behavior to perceived behavioral rules. Others see the social regularities, but just don't change their behavior as a result. Most people fall somewhere in the middle – we may care what others do, but it isn't our only concern. This parameter allows us to investigate these different cases systematically.

The main predictions of the model can be summarized as follows:

1. Norm emergence is incompatible with an adverse intrinsic propensity. If agents strongly prefer doing something else, then they will not follow the proposed behavioral rule. More generally, this means that norm emergence is not guaranteed in the mathematics of the model. Not all behavioral rules become descriptive norms.

2. Other things being equal, the higher the epistemic sensitivity, the more probable that the proposed behavioral rule will in fact become a descriptive norm.

3. The agents' decisions correlate with those agents to whom they assign high expected compliance. Otherwise, decisions are independent of each other.

4. The higher the expected compliance, the more the evidence weighs in favor of or against the hypothesis. In other words, if agents are considered to be highly reliable, their behaviors will have a greater impact on other people than those of less reliable people. This means that it takes fewer more reliable agents for the emergence of a norm than of less reliable agents.

To examine these predictions, we ran a series of simulations to experiment with the effect of the parameters. We present the results in the following subsections. In the next section, we will consider how the model does at providing a general explanatory framework for the phenomena of norm emergence.

3.3.1 Simulation results

The results of the simulations show under which conditions a descriptive norm does or does not emerge, and how the parameters and their interplay affect the final outcome.

In the first group of simulations, we analyzed descriptive norm emergence as a function of the agents' intrinsic propensity. As predicted by Bayes' rule, the higher the priors, the higher the probability that the hypothesis under consideration holds true. In our study, this is reflected by the fact that the probability for descriptive norm emergence increases when intrinsic propensities for the behavioral rule increase in intensity, as shown in Figure 2.2. We see that full convergence on the candidate behavioral rules obtains for high values of intrinsic propensity combined with moderate to high values of expected compliance and epistemic sensitivity. When the intrinsic propensity decreases, the percentage of individuals following the norm decreases proportionally. This is what we expected from the mathematics of the model and expresses the idea that if the preference for the behavior is low then the chance that it spreads are low as well. This is unsurprising – more preferable

behaviors are more likely to spread in a population. Less preferable behaviors have a harder time. We may pay attention to what others do and change our behavior to be more compliant, but we have our limits.

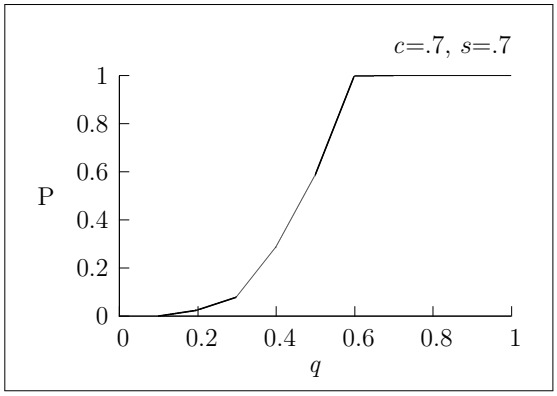

Figure 3.2: Probability P of a norm to emerge as a function of intrinsic propensity q combined with medium expected compliance (c=.7) and medium epistemic sensitivity (s=.7) .

In the second group of simulations, we analyzed the conditions for norm emergence as a function of the expected compliance parameter. By Bayes' rule, the likelihood affects the estimate of a certain hypothesis in such a way that the higher its value, the more significant the evidence is for the hypothesis at stake. In our study, this is reflected by the fact that descriptive norms emerge when reliable individuals follow the behavioral rule and they do not emerge when unreliable individuals follow the behavioral rule. To study the role of the expected compliance parameter on the individuals' decisions over time, we monitored the posterior probability of a few spatially randomized agents, from the beginning of the simulation to the equilibrium point. The simulations results showed that agents tend to follow the norm when the expected compliance parameter is high and it is combined with high intrinsic propensity. By contrast, other simulations showed that agents do not follow the norm when the expected compliance parameter is combined with low intrinsic propensity. In both cases this happens because everyone considers other people's behavior to be highly dependent on the truth or falsity of the hypothesis. Hence, they only follow the behavioral rule if enough other people follow it, and they don't otherwise, suggesting that only in the

former population does the behavioral rule become a descriptive norm.

We find that the expected compliance parameter has a large influence over the model. Note that the tipping point is when the parameter is at 0.5. When the parameter is in the range (0.5, 1], agents count other agents' behavior as significant evidence for or against the hypothesis. As the parameter value trends towards 1, it is harder for a norm to emerge. This seems a bit peculiar at first, but on reflection, it is straightforward. As expected compliance ramps up, any time we see someone not following the behavioral rule, that is considered to be significant adverse evidence. If people are reliable indicators, and someone is not following the behavioral rule, then we surmise that the rule hasn't become a descriptive norm. Evidential standards get more stringent when we perceive the data to be less noisy. Adverse behavior is less likely to be a fluke, and is instead interpreted as evidence that there is no norm. We illustrate this phenomenon in Figure 2.3, where the probability of the norm is given as a function of the intrinsic propensity and the expected compliance parameter.

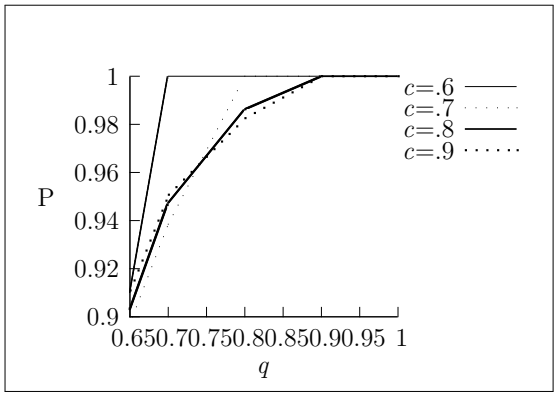

Figure 3.3: Probability P of a norm to emerge as a function of intrinsic propensity q combined with increasing values of expected compliance (c=.6, ..., c=.9 and medium epistemic sensitivity).

Intrinsic propensity is represented on the x axis. The y axis tracks the percentage of agents following the behavioral rule. We plot different values of the expected compliance parameter on the same set of axes to compare them. We see when the expected compliance parameter is lower, agents need to see fewer agents following the behavioral rule to comply with it themselves.

Finally, we used the simulation to explore the effects of the epistemic sensitivity parameter. In the model there is a gap between discovering whether the behavioral rule is a descriptive norm and the decision to follow it. The same probability estimate can induce an epistemically sensitive agent to follow the norm and a less sensitive one not to follow it. Keeping other things equal, we see that when the value of the epistemic sensitivity is low, there is never norm emergence. Social information is simply irrelevant. When instead the epistemic sensitivity is high, at the end of the simulation all agents follow the norm. In other words, for the same probability assigned to the hypothesis that the behavioral rule is a descriptive norm, the difference in epistemic sensitivity induces agents to consider or ignore this information.

These results combine to show the conditions under which Bayesian agents will come to decide that a behavioral rule is in fact a descriptive norm and comply with that behavior. In summary, the simulations help to illustrate to what extent the individual preference matters for the emergence of the norm, and how reliable and socially sensitive individuals affect the process.[15]

Simply relying on a domain-general belief revision mechanism, we can generate rather complex social behavior. As we should expect from our everyday experience with descriptive norms, not every behavioral rule becomes a descriptive norm. Actions that most people don't like much rarely if ever become descriptive norms. When we think people are only moderately reliable in detecting whether a behavior is a descriptive norm or not, it's more likely that we pick out patterns in the noise that might not have been there in the first place. In these results, we find qualitative agreement with results from previous heuristic models, such as that presented in the previous chapter, but as we will discuss in the following section, we claim that the latter approach can offer a deeper explanation of the phenomena.

[15]Several mechanisms might determine the decline of a norm. Two clear candidates are that either a new norm emerges and people switch to it or that an old norm simply fades over time. The former option can serve as a useful description of fashion cycles, while the latter option is particularly clear in the case of fads – eventually they just get old and unexciting. Since we are focusing on norm emergence we leave out considerations of norm decay in order to reduce the complexity of the model.

3.4 Using the Model as a Unifying Explanation

When we consider how to model complex phenomena like norm emergence, there are two different (and complementary) approaches that can be pursued. First, we might want to look at the proximate reasons for norm emergence. Given some assumptions about how we behave in social situations, how do norms come about? Here, we have seen that heuristic models are often going to be particularly useful. Heuristic models present cognitively realistic mechanisms for norm emergence. These heuristic models let us consider the world of boundedly rational agents, and how, despite these limitations, they can systematically create new descriptive norms when the conditions are appropriate. What these heuristic models cannot do, however, is motivate themselves. Nothing internal to the model can tell us why people might track what others do and treat that as evidence for what they themselves should do. Heuristic models – by design – can only speak to proximate causes, not ultimate causes. They do not attempt to ask questions broader than proximate explanation.

Descriptive norms, since they are devoid of any normative force, could seem quite strange if we just look at them with heuristic models. We know that descriptive norms exist, and heuristic models help us understand how they form, given that we're the sorts of agents that look for social rules to follow. But why we are the sorts of agents that have descriptive norms at all goes unanswered. It is only when we turn to a more domain-general style of modeling that we can see that descriptive norms are a side-effect: they are the accidental combination of our system for belief updating and its application to the social realm.

This second approach to modeling complex phenomena offers a deeper explanatory framework. Here, rather than focusing on the details of an individual's thought process, we can ask ourselves if there is a more general explanation for the pattern of behavior. Namely, can we help explain why agents look for rules of behavior to follow in the first place? In particular, is there a way of explaining this phenomenon by demonstrating its connection to mechanisms we understand in other areas of science? Reconstructing the general epistemic process helps us understand the pattern of behavior, even if it does not necessarily capture the precise details of an individual's cognitive processes. What's more, by relying on idealized models of epistemic updat-

ing, such as Bayesian updating, we can more easily see how the phenomenon under study can be understood as instances of a broader framework that has shown success elsewhere. Bayesian reasoning is domain-general, and has been used to examine problems across epistemology and the philosophy of science (Chater and Oaksford 2008; Griffiths et al. 2010; Tentori et al. 2007; Schupbach 2010).[16] Though we may lose a bit in our proximate explanations of individual behavior, by not focusing on the details of the decision processes of cognitively limited agents, our more general reconstruction of this behavior allows us to see how descriptive norm emergence relates to a wider field of epistemic practice.

Bayesian models have seen much success in showing how we can learn about nature, as we come across new evidence (see Jones and Love (2011) for a critical review of the literature). The general process of hypothesis formation, testing, and updating on evidence is well-established in philosophy of science. Even children seem to be successful at learning about causal properties and law-like regularities this way (Gopnik and Tenenbaum 2007; Gopnik et al. 2004). Bayesian belief revision's success in dealing with the natural world provides a reason why we might find individuals naturally extending this apparatus to the social world. Rather than just suppose that people *do* update beliefs based on social information, we can say something about *why* they update beliefs based on social information. In the absence of any norms, there would be no social rules to follow, so no reason to motivate the responsiveness that we have to social cues. However, if we suppose that this responsiveness comes from a domain-general updating mechanism, then we only need to rely on its proven success in other domains.

It is this success in other domains that explains why we would see such a domain-general updating mechanism applied to social cases. Agents are already accustomed to employing such cognitive machinery in a wide variety of cases, and so the social case is just one more instance of using the same basic tools. If anything, it would seem strange to adopt a different epistemic method than the one used so widely in other aspects of one's life. By looking at a wider scope of human activity, we can better see how apparently unrelated tasks can shape our responses in novel situations. The benefit of doing this is that we find that we can get at a more substantive explanation

[16] Again, we would like to note that Bayesian reasoning here is just an exemplar of domain-general reasoning about hypotheses and their evidence. Nothing hinges on Bayesianism in particular. Rather, it is the domain-general belief revision doing the work.

for norm emergence than what is offered in heuristic modeling. In particular, we show that descriptive norms can emerge even if agents are indifferent between a world with no norms and a world full of them.

However, the agents do look for norms. Agents in the model look for norms simply because they see this social situation as just another case where empirical observation can help us uncover law-like regularities. There is no reason to believe that the social world, unlike the natural world, lacks available evidence for how to better navigate it. However, the social situation is unique in that by looking for regularities, regularities are created. If the agents did not try and update their beliefs, and act in accordance with those updated beliefs, norms could never emerge. Unlike private behaviors, norms are public. They are things that we do because others do them. Norms require social expectations. Since descriptive norms have no particular intrinsic value – they don't solve crucial pre-existing coordination problems, they don't improve public morality, and they could easily have been otherwise – they can only come about if enough people believe that they were already there. Once this process begins, norms can start to emerge. In this sense, they are created out of nothing, but become real enough once they come into being.

We can see this clearly in the dynamics of the models themselves. In the initial conditions of the model, there is no norm. People behave based on their intrinsic propensities to act in certain ways. But simply in virtue of *looking* for a pattern in what others do, we start getting a pattern in virtue of more correlated behavior. Once that correlation gets off the ground, the more the agents observe and update, the more they start acting in accordance with the apparent norm. This updating system creates a positive feedback loop. The feedback loop doesn't always start – there aren't always the right conditions for it – but once it does get going, a norm comes to be purely because people were looking for it.

In fact, our system of belief revision will, if anything, overreact to social evidence. In the natural world, when we observe a piece of evidence for, say, whether the moon is waxing or waning, our observation does not affect the moon, nor the observations of others. In the social world, however, if Bob and Carol see Alice change behavior on Monday, and because Bob sees Alice change her behavior, he updates and changes his behavior as well, Carol might now also update on Bob's behavior. But that would just be counting the first piece of information twice. We naturally assume independence

amongst agents in normal situations. Unless we have a good reason for supposing that other people's behaviors or beliefs are correlated, we conceive of other people as making their own decisions. This premise is often true, but can be a catalyst in social phenomena such as norm emergence – the incorrect assumption of independence can lead to large scale behavior change because the first movers have far more influence than people think they have.[17]

Explaining descriptive norm emergence in terms of ultimate, rather proximate explanations also allows us to think more systematically about why we see the norms we do. If we focus on proximate explanations, we run the risk of having to rationally justify each norm. We may get caught up in trying to find ways to argue that each individual norm is utility-enhancing in the same way that the transition from typewriters to word processors enhanced our utilities, rather than the much more defensible claim that the general epistemic processes that have spawned descriptive norms are utility-enhancing. We have argued that the system of belief revision has proven itself to be a massive asset, and so it would naturally be extended to the social realm. It is for this reason that we suspect that epistemic sensitivity might be generalizable. Precisely because epistemic sensitivity in the natural world has proven to be an asset, we can see how its domain might get extended to include the social world as well. When our methods are useful, we try and use them in more places. Our epistemic sensitivity – our disposition to act on the rules we come to discover – can get set by our interaction with the natural world. That we apply it to our social world should come as no surprise if we suppose that we use the same belief revision mechanisms for both. We would need a special reason to think that the dispositions reflected in our epistemic sensitivity ought to be treated differently between our engagement with the natural world and the social world. One way that this might happen is if we come to discover that being disposed to act on rules we find in the social world is harmful in some way. As we have seen, descriptive norms may not be particularly utility-enhancing, but they are also not particularly utility-decreasing. By moving away from proximate explanation, and moving toward ultimate explanation, we can come to understand why descriptive norms emerge. Not because of anything that they do for us, but because they come about from a process that's valuable to us in other areas of our lives.[18]

[17] As discussed earlier, it is in part due to this insight that we chose the network structure that we did for our simulations.

[18] The suggested explanatory framework for the emergence of descriptive norms con-

3.5 Conclusions

As we have seen, once we look at descriptive norms through the lens of a purely epistemological procedure, transformed into a social context, we can see why they can come to be, and persist, even if any particular descriptive norm has no particular value. The process of norm emergence comes along for the ride once we have a general framework for belief revision. As such, we ought to expect (and in fact find) an accumulation of descriptive norms. This itself can reinforce further norm emergence. Once we are in an environment where we are aware that there are already a lot of norms, then it is rational to be all the more vigilant for finding more. This can make us increasingly sensitive to norm discovery. So, while descriptive norms may have started out as a mistaken application of a domain-general belief revision system, their accumulated presence provide a justification for why that belief revision system ought to be applied to them after all. Once our social environment includes descriptive norms as one if its elements, then we have good reason to search for norms as we survey our social world. Descriptive norms may have come to be through a mistake, but their accumulation created a self-justifying reason for their existence.

To conclude, chapters 2 and 3 offer two different but complementary approaches to the emergence of descriptive norms. In both cases, we formulated an agent-based model of the individual decision and studied its features in a computer simulation, in order to observe the effect at the group level. We will now turn to an experimental study on conformity effects in norms compliance. Rather than focusing on the set of descriptive norms, the next chapter will study whether individuals' normative judgment is differently prone to conformity effects according to different types of norm, namely moral, social and decency norms. Unlike descriptive norms, the norms under consideration involve not only empirical expectations but also normative ones, and require an higher-level conceptualization of social organization. The empirical findings constitute the first step towards the formulation of models that look at the effects of the individuals' decisions on the aggregate level.

---

ceives them as a by-product of a Bayesian updating process for detecting regularities in the natural world. In our approach we do not presuppose the existence of descriptive norms, nor we take these norms as input, insofar as we only assume the notion of law-like regularities in nature. We are resting on the idea that the same mechanism that we use to detect regularities in the natural world, when applied to our social world, amplifies the feedback effect and in so doing facilitates the emergence of descriptive norms in society.

Chapter 4

Conformorality: a study on conformity and normative judgment

What is worse, stealing from your neighbor, tipping in Japan or spitting in your glass before drinking?[1] Most people will have no hesitation in answering this question. Perhaps, they may also explain that those behaviors involve different kinds of norms. The first situation seems to concern a moral norm, which holds in all cultures and whose normative force does not depend on people's expectations and preferences. The second involves a social norm, which holds only in particular contexts and whose normative force depends on people's expectations and preferences. The third example, similarly to the first one, involves a type of behavior that is likely to elicit a wave of disgust independent of context or people's preferences and expectations, but just like a social norm, it involves a matter of relatively low seriousness.

This intuitive taxonomy roughly corresponds to a distinction between different kinds of norms, which emerges from the literature on normative judgment in moral psychology (e.g. Bicchieri 2006; Elster 2009; Haidt et al. 1993; Nichols 2002; Turiel 1977). Although there are differences in the way particular researchers individuate different kinds of norms, many would agree that there are features that distinguish moral, social and what can be called 'decency norms.' For example, Turiel (1983, 2002) and his collaborators (e.g. Nucci 2001; Smetana 1993) proposed that people neatly distinguish between

[1] This chapter is based on Lisciandra, Colombo, Nilsenova (under review).

moral norms and social conventions[2] via four main features: (in)dependence of authority, scope, seriousness of violation, and grounds for justification.

According to this distinction, violations of moral norms would be judged as wrong independently of the pronouncements of authorities; moral norms would have universal scope, treated as holding in all places and at all times; violations of moral norms would be judged as seriously bad; and justification of such norms would refer to the harm or injustice suffered by the victim when they are violated. Social conventions, by contrast, would be considered to be authority-dependent, limited in scope, their violations would be less serious than moral violations, and their justification would tend to involve considerations such as the maintenance of social order rather than the harm or injustice suffered by some victim.

It bears emphasis that, for Turiel and collaborators, social conventions, unlike moral norms, are necessarily sustained by general expectations about behavioral uniformities and other people's beliefs. Turiel (1977) makes clear that an assumption informing his work is that "individuals adhere to [social] convention on the bases of (a) the expectation that others do so, and (b) the view that conventional acts are arbitrary" (i.e., there are no intrinsic consequences to the act) (Turiel 1977, p.93). Nucci and Turiel (1978) further explain that "in the case of events that stimulate moral concepts it is not necessary that there be a violation of social regulation for a child to respond to those events as transgressions . . . In contrast, for a child to respond to a social conventional event as a transgression there must be a perceived violation of social regulations or general expectations" (Nucci and Turiel 1978, p.406).

Numerous studies have demonstrated that the distinction between moral norms and social conventions emerges early in human psychology, around three-and-a-half years of age, and is present across different cultures (e.g. Turiel 1983, 2002; Smetana 1993; Nucci 2001). The conclusion that is often drawn in the literature is that moral norms and social conventions, as characterized by Turiel and collaborators, form different kinds of norms, which can be neatly distinguished by human moral psychology (see for a critical discussion Nado et al. 2009).

In agreement with the Turiel's tradition, Bicchieri (2006) distinguishes moral from social norms on the basis of the motivational structure that determines compliance with the norm. While the preference to comply with

[2]In what follows, 'social convention' is used as a synonym for 'social norm'

a social norm is conditional on having expectations about other people's behavior and beliefs, the preference to comply with a moral norm is unconditional. "By their very nature moral norms demand (at least in principle) an unconditional commitment ... Under normal conditions, expectations of other people's conformity to a moral rule are not a good reason to obey it. Nor is it a good reason that others expect me to follow a moral rule" (Bicchieri 2006, pp.20-21). Bicchieri suggests that such an unconditional preference for following moral norms is based on emotional responses that give one independent reasons to comply with the norm (Ibid.).[3]

There have been some criticisms of the distinction between moral and social norm, but we accept them only partially. Recent empirical research has in fact disputed that moral norms and social norms can be neatly distinguished by human moral psychology. Although this research plausibly suggests that the features that allow us to distinguish between different kinds of norms can be more subtle and intricate than what is suggested by Turiel, or by Bicchieri (2006), we maintain that the Turiel tradition and Bicchieri's (2006) are on the right track.

Kelly et al. (2007), for example, had experimental participants to evaluate violations of moral norms that involved harm to others, but in cultures and societies far away in both time and space. Such violations were often judged to be tolerable by Kelly and colleagues' participants. On the basis of their experimental data, Kelly et al. (2007) concluded that skepticism is justified about the association between harm and morality existent in the Turiel tradition. However, Kelly et al.'s interpretation of their data is not free from problems, as shown by further research that Sousa et al. (2009) carried out (see also Sousa et al. 2009; Stich 2009).

Moreover, Nichols (2002) and Haidt (2001) showed that disgusting behaviors may be perceived as seriously bad as moral transgressions, albeit they do not involve harm or injustice to others. According to Nichols (2004), disgusting behaviors might be governed by an idiosyncratic kind of emotionally-laden norms, distinct from moral and social norms, which we call 'decency norms'. We accept that decency norms are distinct kinds of norms. How-

[3]Interestingly, also for Turiel, emotions are prominent aspects of moral norms. Reporting on children's reactions to different norm transgressions, Turiel (1977) writes: "The feedback in the context of moral transgressions generally focused on the effects of actions upon others and on emotional reactions. In contrast, the feedback in the context of social-conventional transgressions focused on aspects of social order, such as rules, sanctions, and norm violations" (Turiel 1977, p.110); see also Nucci and Turiel (1978).

ever, we question that decency norms are moral norms in the way that Nichols (2002) or Haidt (2001) would argue.

Furthermore, judgments about certain types of normative behaviors, but not about others, may well be more resistant to group pressure. Intuitively, given that moral norms are typically assumed to be non-negotiable, we might expect that judgments about, for example, stealing, will be less easily affected by conformity, compared to a judgment about a social norm such as tipping or about a decency norm such as spitting in your glass before drinking. To our knowledge, it has never been experimentally investigated whether different kinds of norms can be distinguished by the degree to which they are affected by peer-group judgment. Answering this question will contribute to progress both in understanding which features allow our mind to selectively distinguish between different kinds of norms, and specifically how social cues impact normative judgment.

In light of previous evidence about the developmental and cultural robustness of moral norms, we hypothesize that the norms that are most resistant to peer-group judgment will be moral norms – as characterized by Turiel and collaborators. Norms that are the least resistant to peer group judgment will be social norms – corresponding to Turiel's conventional norms. With respect to decency norms, if they are found to be significantly different from moral norms in their resistance to conformity effects, then disgust might not be essential to moral judgment, and, at the same time it will probably be insufficient to lead people to morally disapprove of a behavior where no harm or injustice is involved. To test these hypotheses, the present study employed, for the first time in moral psychology, Asch's (1951, 1955) group conditioning paradigm. We compared participants' individual judgments concerning the violation of moral, social, and decency norms, to the judgments the same participants gave in the presence of other people expressing different opinions. Finally, given that nonverbal, social cues such as eye contact, facial expressions and tone of voice seem to play a crucial role in defining in-group social identity and its prototypical (normative) behavior (Hogg and Reid 2006) as well as in facilitating reaching agreement within a group (Hiltz et al. 1986), we hypothesized that the degree of awareness of the other persons – so-called social presence (Short et al. 1976) – in the group conditioning situation might have an effect on conformity. To identify the possible effects of available nonverbal display, we tested whether being unable to see and hear each other results in a lower degree of conformity.

4.1 Test of Material

The test of the experimental material consisted of evaluating thirty scenarios that described a transgression of some norm. The scenarios were based on examples that are found in the philosophical and psychological literature. They included descriptions of behavior involving, for example, some injustice or harm to other people (for what we pre-experimentally took to be moral norms), the infringement of general expectations, or agreements concerning, for example, fairness, reciprocity, or behavioral uniformities that typically regulate interactions between individuals (for what we took to be social norms), and behaviors associated with physical uncleanliness, 'creepy-crawlies' or non-standard sexual practices (for decency norms). The aim was to test if the scenarios would be interpreted by the subjects as instances of moral, social, and decency norms, respectively.

In the test, we also considered the potential impact of personal distance to the perpetrator of the norm transgressions. One could argue that violations that personally involve the participant could trigger emotional processes (Greene et al. 2001, 2004) that might be difficult to evoke with a scenario-based experimental method. If that is the case, we might expect respondents to evaluate differently scenarios concerning strangers (typically employed in moral psychology) to those where the perpetrator is known to the respondent.

4.1.1 Method

Participants. 68 Dutch students (57 female) were recruited from the undergraduate student population at the Tilburg University. They were randomly divided between two conditions and received course credits for their participation.

Design and Instrumentation. The test of the material had a 2x3 mixed design with Distance (scenario concerned a stranger as opposed to a friend/family member) as the between-subject independent variable and Norm Type (moral, social, decency) as the within-subject independent variable. The 30 scenarios were presented in English and described violations of moral, social, and decency norms (ten scenarios per Norm Type, see table 4.1 for examples and table 4.2 for a list of the transgressions employed, classified per type of norm).

The participants were asked to evaluate the scenarios with respect to the

Norm Type	Scenario
Moral	One day Lauren invites Aaron to her place for tea. Aaron accepts even though he doesn't know Lauren very well. They are having their tea, when Lauren has a sexual urge. She wants to have sex with Aaron. Aaron is not willing, he tells Lauren, tries to fend her off, but he can't. Lauren tears off Aaron's clothes and she has sex with him. *On a scale between 1 and 7 how strongly do you approve/disapprove of Lauren having sex with Aaron?*
Social	Michiru, Mauro and Robert are at the pub together. Michiru buys the first round of drinks for everybody. Mauro buys the second. When they have finished their second drink, Robert walks to the bar and buys a drink only for himself. Michiru and Mauro buy their third drink for themselves. *On a scale between 1 and 7 how strongly do you approve/disapprove of Robert buying a drink only for himself?*
Decency	Susan usually has cereals for breakfast. One morning she realizes she finished her favorite cereals. She has only an old pack with grubs and insects inside. She puts them in a bowl and microwaves it first to kill the germs. Then she eats them. *On a scale between 1 and 7 how strongly do you approve/disapprove of Susan eating cereals with insects and grubs for breakfast?*

Table 4.1: Examples of experimental scenarios involving a violation of normative behavior.

Norm Type	**Scenario**
Moral	1. Getting drunk while being the designated driver
	2. Wife cheating on her loving husband
	3. Not paying taxes in Italy
	4. Catching frogs and pouring boiling oil on them
	5. Woman forcing a man to have casual intercourse
	6. Harming the environment to increase profits
	7. Buying a luxury car during famine in Ethiopia
	8. Not voting in EU elections with a low turnout
	9. Keeping slaves 200 years ago
	10. Downloading music from the Internet illegally
Social	11. Having a sexual intercourse in a mosque
	12. Not taking vengeance for one's sister on Corsica
	13. Coming to a dinner without a gift for the hosts
	14. Enjoying rounds of drinks but not contributing
	15. Not leaving a tip in a restaurant in the U.S.
	16. Playing cards in a church during a funeral
	17. Not sharing gained money during a game
	18. Making a phone call in a cinema
	19. Playing further after an opponent has been injured in a game
	20. Leaving a shopping cart in the line to shop further
Decency	21. Eating parts of the deceased relatives' bodies
	22. Wearing a sweater that once belonged to Hitler
	23. Brother and sister making love
	24. Eating one's dog after it was killed by a car
	25. Eating cereals with insects for breakfast
	26. Sexual partners urinating on each other
	27. Bathing in chicken blood
	28. Sheep ranchers having sex with sheep
	29. Growing worms in the bedroom and eating them
	30. Spitting in glasses before drinking
	Note: * p<.05, ** p<.01, *** p<.001

Table 4.2: Violations involved in the scenarios classified according to the Type of Norm.

following four items, each operationalized in terms of a 7-point scale anchored at the ends with (1) *strongly disagree* and (7) *strongly agree*: *Badness* ('X's behavior is very bad'), *Disgust* ('What X did is nauseating'), *Time/Place* ('In a different time/place, it is OK to do what X did') and *Authority* ('If the law allows it, it is OK to do what X did'). These items were based on the properties identified by Turiel (1977), Kelly et al. (2007) and Nichols (2002) as characteristic features of the different types of norms. In the Stranger condition, the scenarios concerned unknown individuals with invented names; in the Friend/Family condition, the names were replaced with phrases such as 'your room-mate', 'your best friend' or 'your parents'.

Procedure. The test was administered online and presented as a study of Dutch taboo subjects. The participants were invited to read each scenario as if it were describing a situation that actually happened.

4.1.2 Results

We analyzed the results with mixed ANOVAs with Norm Type and Distance as independent variables and the score on each of the four items as the dependent variable. The data showed no significant main effects of Distance for *Badness*, $F(1,66) = 2.945$, $p = .091$, for *Disgust*, $F(1,66)<1$, $p = .579$, for *Time/Place*, $F(1,66)<1$, $p = .620$, and for *Authority*, $F(1,66)<1$, $p = .521$. There were also no significant interaction effects between the variables Norm Type and Distance for *Disgust*, $F(2,132)<1$, $p = .430$, for *Time/Place*, $F(2,132) = 2.850$, $p = .061$, and for *Authority*, $F(2,132) = 1.959$, $p = .145$. There was an interaction effect between Norm Type and Distance for *Badness*, $F(2,132) = 4.527$, $p = .013$, $\eta_p^2 = .06$.

These results indicate that scenarios that involved the participants' friends and family members were not judged differently than the scenarios involving strangers. The scales evaluated for each scenario distinguished between the three Norm Types as summarized in table 4.3.

Item	Moral	*SD*	Social	*SD*	Decency	*SD*	η_p^2
Badness	5.3	0.60	4.5	0.72	5.0	1.10	.30*
Disgust	4.9	0.73	3.7	0.96	5.9	0.79	.84*
Time/Place	2.9	0.72	3.7	0.86	3.3	1.10	.28*
Authority	2.9	0.61	3.5	0.81	2.9	1.00	.21*

*Note.** $p<.05$

Table 4.3: Summary of the mean participants' judgments in the two survey conditions per item, measured on a 7-point disagree/agree- scale (N=64).

For the property *Badness*, *Decency* and *Time/Place*, the three types of norms differed significantly from each other. The perception of *Badness*, $F(2,126) = 25.161$, $p < .001$, $\eta_p^2 = .29$, differed form moral norms compared to decency norm ($p = .008$) and social norms ($p < .001$), as well as for decency norms compared to social norms ($p = .004$).

With respect to Disgust, $F(2,126) = 174.631$, $p < .001$, $\eta_p^2 = .74$, all the norms differed from each other with $p < .001$. *Time/Place*, $F(2,126) = 15.430$, $p < .001$, $\eta_p^2 = .20$, could distinguish between moral and decency norms ($p = .006$) and social norms ($p < .001$), but not between decency and social norms ($p = .117$). For *Authority*, a pairwise comparison showed a difference between moral and social violations ($p < .001$), and decency and social violations ($p = .001$), but no significant difference between moral and decency violations ($p = .870$).

Finally, we inspected the correlations between the scores assigned to scenarios within a Norm Type, focusing on the properties that in the literature are assumed to be relevant for distinguishing between the norms, to wit *Badness* for moral norms, *Disgust* for decency norms, and *Time/Place* and *Authority* for social norms. The analysis showed no outliers within the categories, i.e., scenarios that would be negatively correlated with other scenarios in the category with respect to the distinguishing property. The Cronbach's alpha coefficients (measures of internal consistency of the scales) were $\alpha = .62$ for moral violations on the *Badness* scale, $\alpha = .73$ for *Disgust*, and $\alpha = .60$ for *Time/Place* and $\alpha = .56$ for *Authority*, showing the highest internal consistency with respect to judgments of *Badness* and *Disgust*. In the case of decency violations, the Cronbach's alpha coefficient was relatively high on

all the four scales, with = α .81 on the *Disgust*-scale, α = .85 for *Badness*, α = .85 for Time/Place and α = .84 for *Authority*. For social norm violations, there was an acceptable internal consistency for all the four scales, with α = .72 on the *Time/Place*-scale, α = .75 on the *Authority*-scale, α = .80 on *Disgust* and α = .68 on *Badness*.

4.1.3 Discussion

The results of the material test show that the characteristic properties of three types of norm violations, which have been identified in the literature (the seriousness of the violation, its dependence on time/place and on an authority, and the feeling of disgust it evokes) distinguish between the scenario types employed in the test and thus validate the original classification of the scenarios, which was based on the literature. The participants were not more sensitive to scenarios involving a familiar person compared to those concerning a stranger and the distinction was not taken into consideration in the subsequent experiment, in which we employed the thirty scenarios from the material test.[4]

4.2 Experiment

4.2.1 Method

Participants. 97 Dutch native speakers (66 female), all with a good command of English, between the ages of 19 and 49, were recruited from the undergraduate student population at Tilburg University and received course credit for their participation.

Design and Instrumentation. The experiment had a mixed 3x3 design, with Norm Type (moral, social, decency) as the within-subject variable and Social Presence (high, low and control) as the between-subject variable. The questionnaire consisted of the thirty short scenarios described above and 10

[4]The first part of the study has been conducted in order to get independent evidence about the norm taxonomy employed in the actual experiment. In this way, it has been shown that the distinction between social, moral and decency norms is not only based on intuitive criteria, but –more carefully– on certain characteristic features of each group of norms. Another viable option could be to run a cluster analysis in order to observe whether the clustering of the scenarios on the basis of their sensitivity to group pressure corresponds to the initial classification. Thanks to Jason Alexander and Jan-Willem Romeijn for pointing this out to me.

distractors. The distractor items had content similar to the experimental items in that they involved different kinds of norm violations.

The participant's judgment was measured on a 7-point scale anchored at the ends with (1) *'strongly disapprove'* and (7) *'strongly approve'*, with participants indicating their acceptability judgment for each scenario, first in an individually completed online questionnaire and, two weeks later, in a group condition with three confederates. In the online version of the questionnaire, participants were also asked to indicate for each scenario if they were certain of their judgment (*yes/no*).

For the thirty experimental items, the confederates' answers employed in the group condition were chosen using the mean of the participants' answers in the first measurement, with two scale points added to the mean in the 'least desired direction'. For each item, the 'least desired direction' was operationalized on the basis of which half of the scale (i.e. either the 'disapprove' or 'approve' half) the participants used less often in the individual condition. The confederate answers were unanimous on the thirty experimental items and differed for the ten distractor items. In the control condition, participants merely filled out the online questionnaire twice with a two-week period in between. For the first measurement in the individual condition, we used two sequences of the online questionnaire to test for possible order effects. In the second sequence, the questions were presented in reverse order.

Procedure. In the group condition with high social presence, the participants were seated together with three confederates and they could see each other's expressions and hear each other's voice. In total, 24 students, both male and female, acted as confederates. The experimental leader (a female for half of the trials and a man for the other half) read each scenario and the participants gave their answers in the order: confederate 1 - confederate 2 - participant - confederate 3. The participants were informed that the answers they gave online were lost due to a server error and had to be collected again. In order to avoid differences in cognitive load between the first and the second measurement, the participants were supplied with the text of the scenarios on paper.

In the condition with low social presence, the participants were seated in front of a computer screen in the same room as the confederates but could not see their faces. In order to exclude vocal cues, they all indicated their judgments for each scenario by selecting their answer on the screen, where

both the scenarios and the answers of the confederates were presented. At the end of each session, the participants were interviewed and debriefed. None of the participants reported having difficulties in judging the scenarios.

4.2.2 Results

A Mann-Whitney test of judgments per scenario collected during the first measurement revealed no effect of presentation order on participants' judgments. The data from the first measurement in all three conditions, summarized as the mean value of the participants' judgments per scenario, were used to examine the homogeneity of variance for the three types of norms. The Levene Statistic showed that the assumption of equal variances was valid, indicating no systematic differences in answer distributions.

In order to test if all three types of scenarios were judged with the same certainty, we first compared the categorical data indicating participants' certainty of their approval judgments. There was no significant difference between the three scenario types, $\chi^2(2) = .16$, $p = .920$; for most scenarios (92.7%), the participants indicated themselves to be certain of their judgment.

In the subsequent analyses comparing the first and the second measurement, we excluded cases where the participant had the same judgment during the first measurement as the confederates in the group condition (13% of the total of 2910 experimental trials, distributed equally over the three Norm Types).

We calculated Conformity (C) using the approval judgments given by the participants in the individual (M_1) and the group condition (M_2) and the confederates' opinion (O), as $C = |O - M_1| - |O - M_2|$. A positive value of C represents instances where the participant's judgment shifted closer to the confederates' opinion, a negative number stands for cases where the distance increased and 0 for cases where the distance remained the same.

Given that the dependent variable Conformity was not normally distributed (Shapiro Wilk's test < 0.5), we used nonparametric tests throughout. We first examined whether male and female participants differed in their overall Conformity scores in the two conditions involving confederates. A Mann-Whitney U showed no significant effect for gender (U = 374.00, z = -1.43, p +.154). We used the Kruskal Wallis test to analyze the difference between the experimental conditions with high and low social presence

and the control condition. A Mann-Whitney test with Bonferroni correction showed that the condition with high social presence differed from the Control condition for all three Norm Types, as well as the Total Conformity.

The condition with low social presence differed from the Control condition in the case of Social Conformity and the Total Conformity, but not for Moral and Decency Conformity (see table 4.4).

	High SP			Low SP		
Conformity	*U*	*p*	*r*	*U*	*p*	*r*
Moral	297.000	.01	-.328	397.000	.13	-.188
Social	88.000	.00	-.701	297.500	.00	-.356
Decency	184.000	.00	-.529	464.000	.56	-.074
Total	116.000	.00	-.650	323.500	.01	-.311

Note. $df = 2$. *SP* = Social Presence

Table 4.4: Mann-Whitney tests for the conditions with high social presence (N=33) and with low social presence (N=35) compared to the Control condition (N=29).

The medians for the three types of norms in the three sets of conditions are reported in table 4.5.

In order to examine the difference between the three Norm Types (moral, social, and decency) in detail, we used the Friedman test to compare the level of Conformity separately in the two experimental conditions, with high and low social presence. The analysis showed that the three Norm Types differed only in the condition with high social presence ($\chi^2(2) = 7.09$, $p < .05$), but not in the condition with low social presence ($\chi^2(2) = 2.97$, $p = .227$) - see table 4.5. In the condition with high social presence, participants conformed the most to the scenarios describing social violations ($Mdn = .600$), compared to decency violations ($Mdn = .546$) and moral violations ($Mdn = .400$). Wilcoxon tests with the Bonferroni correction (effects reported at a .0167 level of significance) showed that Conformity to judgments of moral violations differed from Conformity to social ($p = .003$, $r = -.471$) and decency violations ($p = .008$, $r = -.417$), but Conformity to judgments of social violations did not significantly differ from Conformity to decency violations ($p = .187$, $r = -.160$).

Finally, we ran a secondary analysis of the consistency of answers across

	Condition			Statics	
Conformity	*HighSP*	*LowSP*	*Control*	X^2	p
Moral	.40	.20	.00	7.100	.03
Social	.60	.20	.00	33.998	.00
Decency	.55	.00	.10	21.065	.00
Total	.52	.16	.00	30.835	.00

Note. $df = 2$. SP = Social Presence

Table 4.5: Median Conformity differences in the three experimental conditions (low social presence, high social presence and control) by Norm Type (N=97). The scores express the change in distance to the confederate's opinion, higher score indicating higher conformity (0 = no change).

measurements, calculated as the absolute difference between the participant's first and second measurement (independent of the confederates' answers). The results showed that, similarly to the Conformity measure, the stability of answers was higher for moral scenarios compared to the other two types; Norm Type: $F(2,\ 188) = 9.95$, $p < .001$, $\eta^2(2)\text{p} = .10$; Condition: $F(2,\ 94) = 6.24$, $p = .003$, $\eta^2(2)\text{p} = .12$; Norm Type * Condition n.s. A pairwise comparison analysis showed a significant difference between moral and social, and moral and decency norms, but no difference between social and decency norms (see Table 4.6 for means and standard deviations).[5]

[5]Also, a further analysis has been conducted which showed that no memory effects could explain the higher stability of moral judgments as compared to judgments of other kinds of norm.

		Condition					
	N	With SP		Without SP		Control	
Norm Type		*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Moral	10	0.81	0.38	0.78	0.32	0.63	0.34
Social	10	1.08	0.41	0.92	0.43	0.73	0.31
Decency	10	1.06	0.50	0.82	0.31	0.76	0.43

Note. $df = 2$. *SP*= Social Presence

Table 4.6: Median Conformity differences in the three experimental conditions (low social presence, high social presence and control) by Norm Type (N=97). The scores express the change in distance to the confederate's opinion, higher score indicating higher conformity (0 = no change)

4.3 General Discussion

Earlier research in psychology has examined, on the one hand, the effects of authority on obedience and norm compliance (Milgram 1963), in-group/out-group effects on moral behavior (Tajfel 1981), and the consequences of emotional cues on people's normative judgments (Schelling 1978; Wheatley and Haidt 2005). On the other hand, research studies on humans and nonhuman primates have shown that both species tend to adjust their behavior and beliefs toward others in their social circles (Cialdini and Goldstein 2004; Whiten et al. 2005). In humans, conformity can affect judgments ranging from perceptual line-length estimates Asch (1951) to more complex behaviors, such as energy saving (Schultz et al. 2007).

Combining both threads of research on normative judgment and conformity effects in an original way, our experiment focused on understanding the effects of peer pressure on individuals' normative judgments. The results of our experiment indicate that while all normative judgments tend to be affected by peer-group judgment to some degree, the effect is the strongest for social and decency norms, which are most likely to be influenced by peer-group conditioning. Moreover, the effect is especially pronounced in situations involving a higher degree of awareness of others, operationalized

in terms of the availability of nonverbal display. The degree of conformity to other people's normative judgment as such can then be used to independently motivate the distinction between moral norms and social norms proposed by Turiel and collaborators and by Bicchieri. Our findings are congruent with previous research both on conformity effects in computer-mediated communication (Smilowitz et al. 1988; Bordia 1997; Cinnirella and Green 2007; Laporte et al. 2010), as well as with studies conducted by Bicchieri (2008) and Cialdini et al. (1991) on the effects of expectations about other people's compliance with a norm.

To explain our main results, it can be suggested that the predisposition we have towards conformism to common behaviors and shared opinions of our own group is counterbalanced by the robust influence that a specific kind of norm, that is moral norms, has on our mind. Hence, the degree of dependency on other people's judgments makes it possible to reliably distinguish moral norms from different types of norms. On this basis, it can be suggested that moral norms constitute a natural kind in human moral psychology.

Furthermore, the fact that decency norms appear to be less stable than moral norms lends support to critical reviews according to which there is weak evidence that disgust is a moralizing emotion (Huebner et al. 2009). Although disgust may be implicated in moral judgment, it is probably neither sufficient nor necessary for moralization to occur (Royzman 2009). A number of variables, including group size, group composition in terms of gender and age, as well as cultural background of the participants may influence the outcome of group conditioning experiments and should be explored in future studies of conformity to judgments of norms. However, if human psychology is selectively sensitive to recognize and implement moral norms, which might constitute a cognitive domain robust to conformity effects, then our main result should be found across different groups and cultures.

One important issue for future research is that a more fine-grained analysis of the content of the scenarios used in our study is necessary in order to make firmer, and more specific claims about the psychological nature of distinct kinds of norms. Some of the items we used might be revised so as to enrich them with more context, which may be relevant to judge the kind of transgression involved. For example, privacy and prudential considerations that a decency scenario might activate are relevant to make firmer conclusions about decency norms. With respect to privacy, if some of the transgressions of decency norms were interpreted as being done in the pres-

ence of other people, they would involve offense, which can be considered as a specific type of harm. This may make some of the decency scenarios not that different from the moral ones (cf. Royzman 2011). With respect to the prudential issue, some of the decency scenarios might have been interpreted prudentially, in terms of the unhealthy consequences for the perpetrator, rather than disgusting practices.

The language of the experiment might also be a factor; in our study, we presented English scenarios to Dutch participants. Even though their knowledge of English was good, the fact that they were evaluating norm transgressions in a non-native language may have reduced the impact of our manipulation (Puntoni et al. 2008). Arguably, this might affect decency norms more than moral ones.

Additional research is also needed to validate the scenario-based technique employed here by linking it to behavioral data collected in natural and simulated (game) settings (van Lankveld et al. 2011), possibly using methodology that has been previously employed to determine personality profiles.

Chapter 5

Towards a methodological account of robustness analysis

5.1 Introduction

In the previous chapters, a set of models for the emergence of norms and a series of experiments on norms compliance have been presented and the robustness of their results investigated. The method of testing whether the same predictions follow from a set of different theoretical or experimental assumptions is known as robustness analysis. Correspondingly, the predictions of a model or an experiment are said to be robust if they hold true even when some of the assumptions, from which they are derived, have been challenged and replaced by others.

Whereas in the experimental sciences robustness analysis is used as a test of the effect of possible confounders on the empirical results, the arguments in support of robustness analysis in non-experimental contexts are often left implicit or are unreflectively imported from the experimental sciences. This final chapter will be dedicated to an examination of the logic behind this practice as it is used in theoretical models.

Intuitively, the general idea behind robustness analysis is as follows: Suppose that we have a model, based on a number of initial assumptions, from which a number of predictions are derived. If the initial assumptions are simplified representations of the real-world phenomenon, it is natural to ask how the predictions of the model can apply to the real-world phenomenon, where such simplifications do not hold. One strategy is to replace the initial assumptions with different ones, to observe whether the predictions hold true

across conditions. Consistency of the results would suggest that the unrealistic assumptions were irrelevant to the final result; inconsistency would show that the predictions were not independent of the specific initial assumptions.

A classic example of robust explanation is Schelling's segregation model (Schelling 1978). This model describes the dynamics that lead to racial segregation within social groups. More generally, it applies to any situation where individuals have preferences that tend to generate social clusters, i.e different tastes, language, social status, sex, age, etc. Schelling's model can be represented by means of a checkerboard, with dimes and pennies, standing respectively for a certain metropolitan area and for the individuals of two different groups, for example Blacks and Whites. The behavior of the individuals is determined by a decisional rule that makes them move from one place to another, until the composition of the neighborhood meets their preferences. As it turns out, regardless of their initial distribution in the metropolitan area, Black and White citizens will end up being segregated in two different parts of the city, as a consequence of their preference for having at least half their neighbors of their own color. Interestingly, the model predicts segregation not as a consequence of the preference of the individuals for segregation itself, but as a by-product of their preference for having a few neighbors of the same ethnicity. With respect to robustness, the fact that segregation occurs across different initial positions is considered to be a virtue of the model, as it suggests that the result does not depend on one specific assumption, i.e. a simplified representation of the distribution of the individuals in space.

The robustness of Schelling's model has been tested under a number of different assumptions, other than initial position. For example, Bruch and Mare (1989) have shown that segregation occurs under different updating rules, structures of neighborhood, and alternative choice functions. Muldoon et al. (2012) have shown that segregation takes place even when the individuals prefer to be in the minority group of their neighborhood. Overall, these studies are meant to establish whether the same effect follows under more plausible assumptions than the original ones, such as more fine-grained utility functions or less stylized metropolitan areas. In this respect, the relation identified by one model is more robust than another, to the extent that it is resistant to a larger set of variations in the underlying features of the system being investigated.

In the philosophy of science literature, robustness analysis of scientific

models has been either defended or rejected as a confirmatory device and the core of the dispute is about whether this practice can guide the comparison between a model and the empirical world. The critics consider the analysis to be an *a priori* method of inquiry, which only assesses the effects of variations in the assumptions of a model on a theoretical level. Its advocates value it as an effective way of increasing confidence in the theoretical predictions, before or when it is not possible to test them against the empirical data. More generally, the epistemological problem of robustness analysis concerns what we can conclude from the stability of our predictions through changes in the assumptions from which they are derived.

In a recent paper, Woodward (2006) distinguishes various senses in which the notion of robustness has been used across scientific areas and claims that each one has its own criteria of justification. The major distinction he draws is between experimental robustness and theoretical robustness. The former refers to the stability of a certain result across multiple experimental techniques, or variations of the same experimental setting, where the consistency of the measurement outcomes is taken to confirm the initial hypothesis. The latter applies to theoretical models and investigates whether the same predictions can be derived from a set of different assumptions.[1] Within the theoretical domain, a further distinction has been drawn by Weisberg and Reisman (2008) between:

1. *parameter robustness*, which refers to variations in the initial conditions or in the values assigned to the parameters of the model;

2. *structural robustness*, which refers to changes in the parameters included in the model;

3. *representational robustness*, which refers to modifications in the formal structure in which the model has been implemented.

Throughout this chapter, it will be argued that even within the context of theoretical robustness, different criteria of justifications apply across differ-

[1]This is however a broad distinction, which does not take account of the variety of uses of robustness analysis across scientific domains. For example, in the study of complex systems, robustness analysis is combined with sensitivity analysis as a method of quantifying the effect of uncertainty at the level of the parameters on the final predictions. In statistics, robust estimators are those unaffected by outliers in the data. More on this can be found in the literature on robustness in econometrics (Leamer 1983) and climate sciences (Parker 2011; Pirtle 2010).

ent cases. More specifically, the motivations for robustness analysis as a method of proving that a certain result is robust, with respect to different assumptions about the system being investigated, are different from those related to different ways of modeling the same component within a model. The former, which concerns the first two senses of robustness in Weisberg's classification, is a method of observing whether and how the introduction of new ingredients into a model affects its predictions. The latter, which corresponds to Weisberg's third category, is a method of observing whether and how different ways of expressing the same ingredient affects the predictions. In the literature, however, the difference between the criteria behind these two senses of robustness has not been sufficiently appreciated and this has generated misunderstandings about the nature and scope of robustness analysis. The justifications in favor of one approach are not by default relevant to the other and one sense of robustness analysis will not be weakened if the other is shown to be untenable.

This chapter is organized as follows. In Sec.2, I start by delineating the concept of robustness analysis and briefly present the main arguments for and against this method. I suggest that the lack of agreement in the scientific community about the epistemological status of robustness analysis depends partly on the absence of a unified account of robustness analysis across domains. In Sec.3, I argue in support of robustness analysis as a means of testing the role of the assumptions of a model, with an example from population biology, and I illustrate the relevance of this practice in the field. In Sec.4, I present a case study from economic geography, where robustness analysis is considered to be a method of testing whether the same result can be derived from different *tractability assumptions*, namely different mathematical formulations of the same factor. I conclude by pointing out a number of difficulties that emerge from the robustness analysis of tractability assumptions. I claim that the objections to robustness analysis in the latter case do not undermine the previous ones (Sec.5).

5.2 For and Against Robustness Analysis

In the experimental sciences, the robustness of a certain result – with respect to changes in the experimental setup – is considered to be a way of testing the result achieved in the laboratory and of ruling out the effects of possible confounders (see e.g. Guala 2005; Guala and Mittone 2005; Hacking 1983).[2] In the case of non-experimental science, the analogous question is whether the robustness of the model's predictions – with respect to changes in the initial assumptions – is a means of confirming the theoretical predictions. The two tests proceed on similar lines: while doing experiments, scientists modify the experimental setting in some aspects in order to observe whether relevant effects follow; in the case of models, scientists modify the theoretical structure and analyze the consequences of this change.[3]

Robustness analysis, in experimental contexts, is usually not considered to be a test of the external validity of the phenomenon under scrutiny, but mainly of the effect observed in the laboratory. By contrast, it has been contended that – in the case of scientific models – where there is no such distinction between the experimental setup and the world, robustness analysis can provide support for the theoretical predictions of the models. The rationale behind this claim is that robustness analysis offers a means of addressing the problem of the unrealistic assumptions of theoretical models, which is the problem of how scientific models can represent the empirical world, despite being based on idealizations and abstractions that do not literally match with the distinguishing features of the phenomenon under consideration (Frigg and Hartmann 2012). In physics, for example, the motion of a simple pendulum is explained assuming a uniform gravitational force, even though there is no

[2] A classic example where robustness analysis has proved to be successful in an experimental context is Perrin's determination of Avogadro's number, which in turn was considered to be crucial to assess the existence of molecules (Perrin 1923). The consistency of the result through different and independent methods of measurement was decisive in ruling out the possibility that that result was the effect of one specific measurement tool. In this chapter, I distinguish – within the realm of the empirical sciences – between experimental robustness, as a method of testing inferential relations through changes in the assumptions of the experimental setup, and measurement robustness, as a method of measuring the properties of physical entities through different measurement tools.

[3] See Guala (2005, pp. 224-229) for a detailed distinction between the notion of robustness analysis and external validity.

such thing as uniform gravitational force. In the model, the earth is represented as a perfectly homogeneous sphere, there are no other celestial bodies that exert a gravitational influence on the pendulum and no other kinds of forces, apart from the gravitational ones, that affect its motion. This way of proceeding is not necessarily detrimental to the predictions of the model. If it can be shown that the same predictions follow from different assumptions, corresponding to different degrees of proximity to the real phenomenon, then these predictions are not necessarily undermined by the initial unrealistic assumptions.

This sort of analysis, however, is beset with difficulties. Part of the problem is to ascertain whether the role of the assumptions in a theoretical model is as simple as stated above. Whereas, in the case of the pendulum, the role of the omitted factors is negligible, in other circumstances it is not so obvious whether the simplifying assumptions are not excluding relevant features of the system under analysis. More generally, the critics of robustness analysis (Cartwright 1991; Odenbaugh and Alexandrova 2011; Orzack and Sober 1993; Sugden 2001) raise a number of objections to the claim that this method can boost confidence in an hypothesis. First, they maintain that robustness analysis is a non-experimental method of inquiry, at odds with the fundamental principles of scientific method, according to which our hypotheses should be tested against the empirical evidence rather than against *a priori* reasoning. Examining the role of assumptions by substituting them with new ones is only a way of remaining in the theoretical sphere of a model.

Further reasons to be skeptical of robustness analysis are that varying assumptions might all lead to the same result, which is itself wrong: it is not sufficient that the outcomes of different analyses are consistent with each other for them to be true. In this regard, Orzack and Sober discuss the following case: "Consider, for example, all models in which natural selection is said to be the only force acting on a population. This assumption has a consequence that population size is infinite. Accordingly, this is a robust prediction for this set of models. This gives us no reason, however, to think that populations in nature really are infinite." (Orzack and Sober 1993, p.538)

Yet another reason for skepticism is that lack of robustness does not necessarily imply the falsity of an hypothesis. For example, a certain economic model whose predictions turn out to be accurate when applied to a specific geographic region might fail with respect to another. In such a situation, it is not clear why the fragility of the result should invalidate the model in the

first place. The inconsistency could be taken as proof, either of the fact that the model does not capture the general dynamic of the economy, or of the fact that the market is simply different in the two contexts. More generally, if a certain hypothesis is shown to be sensitive to changes in its theoretical structure, it does not follow that the hypothesis is to be rejected, since its fragility might reflect a genuine feature of the system (Hoover 2006).

Finally, the critics of robustness analysis ask for a more accurate specification of the relation between a certain result and the set of assumptions from which it is derived (Aldrich 1989). Do the assumptions have to cover all possible configurations of the system under scrutiny, i.e. are they mutually exclusive and exhaustive? What if the phenomenon is found to follow from them only with a certain probability? For example, suppose that Schelling's segregation model were to predict that separation by color would occur only according to a number of spatial configurations, i.e. only with a certain probability. How high should this probability be in order for the phenomenon to be considered robust?

Against these objections, the advocates of robustness analysis (Kuorikoski et al. 2010, 2012; Weisberg 2006; Weisberg and Reisman 2008) defend it as an effective guide to scientific research. Consider again the segregation model: if the process under consideration is shown to be independent of a number of specifications of the system under scrutiny, then scientists can remain agnostic about the details of the problem without this undermining the result. This turns out to be a crucial feature in all those areas of research where scientists cannot know the exact configuration of the system they intend to explain. For example, in evolutionary game theory, a standard objection to the validity of certain results about the emergence of cooperative behaviors in human societies concerns their lack of robustness with respect to the individuals' cognitive constraints (Skyrms 1996; Sugden 1986; D'Arms 1998). A limitation on the kind of possible strategies that can be transmitted across generations is the cognitive load they impose on individuals; thus, a result will not be considered significant if it does not stand a test of robustness that takes these limitations into account.

More generally, the partisans of robustness analysis maintain that, even if the method does not make it possible to derive the occurrence of a certain phenomenon deductively, this is not a reason to reject it. The situation is the same as in science in general: inductive inferences always require an inferential leap. In the case of robustness analysis the question is whether a robust

result confirms an hypothesis more strongly than a non-robust one. Certainly, they argue, we do not want to accept a practice that is not epistemically justified, but at the same time we do not want to abandon a methodology that might increase confidence in our hypotheses. On this analysis, the criteria for robustness – such as, that the underlying assumptions must be exhaustive and mutually exclusive – could even be too strict, and lead to the dismissal of certain results that it would be more reasonable to accept (on this argument, cf. Woodward 2006).

Overall, the divergence of views between the advocates and the critics of robustness analysis appears very radical: either scientists are operating on the basis of non-justified procedures, or the reasons in support of these procedures have not been sufficiently elucidated. So far, the debate between the opposing sides has not led to conclusive answers. Part of the difficulty in finding agreement lies in the fact that robustness analysis has been used across disciplines, in each of them in conformity with its own characteristic methods. Applying arguments that are appropriate for a certain context to a different one may have made the criteria that justify this method unclear. In the next sections, different examples of robustness analysis will be provided with the goal of highlighting the distinctive criteria behind them.

5.3 A Case Study of Robustness from Population Biology

An example that helps illustrate how robustness analysis works in biology and related disciplines is provided by Weisberg (2006, 2007), and Weisberg and Reisman (2008) with the Lotka-Volterra predator-prey model. This is a mathematical model which analyzes the dynamics between prey and predators with respect to the size of each population. The Lotka-Volterra model is based on two coupled differential equations and one property derived from them is the Volterra principle. This shows that the introduction of an external cause of death in the system, such as a pesticide, affecting both the prey and the predator, determines a lower decrease in the growth rate of the prey than in that of the predator.

The Lotka-Volterra model relies on a number of assumptions: for example that there is no scarcity of food in the environment. This means that population density is not considered initially as one of the factors influencing the

growth rate of each population, even though it is well known that it usually affects the process. The reason for omitting this factor is mainly to focus on the relation between the prey and the predators rather than on other external features. After deriving the properties of the model in its simplified version, a test of robustness is made by introducing a new assumption representing the carrying capacity of the environment. The Volterra principle still holds once the population density factor is added to the model. Moreover, the principle also holds under other realistic features, which were missing from the original formulation, such as a limit in the number of prey for each predator (see Weisberg and Reisman 2008). Other properties, however, are shown to be sensitive to the change: for example, whereas originally the oscillations representing the abundance of the two groups in their ecological system were stable, this was no longer the case after the introduction of the population density factor. Because the Volterra principle holds under a more realistic assumption, and other features of the baseline model do not, then – Weisberg argues – we have more reason to accept the robust principle than the non-robust properties.

Weisberg also points out that the robustness of the Volterra principle can be tested within different theoretical frameworks, i.e. not only by means of analytic proofs but also in the context of agent-based models (Weisberg and Reisman 2008). In this case, the way of proceeding in order to verify whether the principle is independent of population density is to assign different values to this factor and to observe whether the final result is affected by this change. Whether by computer simulations or formal analysis, the point of the procedure is to test whether a certain relation holds when modifying the assumptions about some features of the environment, which characterize the phenomenon under scrutiny, in order to see whether they affect the dynamics of the phenomenon under study.

One question that emerges from this study is to what extent it is in the hands of the modeler to formulate the new variables in such a way that they will affect (or not) the previous results. The answer to this question, however, is not specifically directed to robustness analysis, but involves the practice of model building in general. It is part of the elaboration of a model to assess whether the predictions are robust to changes in its unrealistic assumptions. Changes in the assumptions may have interesting effects, as may the variables that were initially considered to be causally relevant to the phenomenon under consideration. Just as it is part of experimental practice to check whether

a certain effect is determined by the specific conditions of the experimental setup, it is part of theoretical practice to check whether the predictions depend on the idealized setup of the model. In both cases, in order to test the experimental and the theoretical predictions, the burden of proof lies in the empirical investigation of real-world phenomena, but robustness analysis is a preliminary to empirical investigation. Even when theoretical models are not built with the intention or the possibility that they will be tested experimentally, robustness analysis is inherent in the formulation of the model in the first place, as it provides an indication of how relevant variables might affect the predictions.

In this respect, the Lotka-Volterra prey-predator model offers an example of robustness analysis as a method of finding out whether the theoretical predictions are affected or not by varying assumptions about the system in which the phenomenon under investigation takes place. In a different way, robustness analysis has also been deployed as a method of observing whether a certain result is insensitive to different tractability assumptions, namely different mathematical formulations of the same factor in a model. In the next section, an example of this kind of analysis will be presented and discussed.

5.4 Robustness Analysis of Tractability Assumptions

Weisberg and Reisman (2008) have introduced the term *representational robustness* to describe a test of the consistency of predictions across different mathematical approaches. For example, in the Lotka-Volterra model, the same phenomenon has been analyzed both by differential equations and agent-based models. Similarly, Colyvan and Ginzburg (2003) have suggested that the predator-prey model should be treated by second-order differential equations rather than by first-order differential equations, since the former provides a better description of the change in the population's growth rate.

When considering whether to adopt a certain mathematical structure, or a specific mathematical assumption within the same model, the purpose is the same, i.e. to find an effective combination of mathematical tractability and expressive power. For instance, the degree of specificity achievable via agent-based models is higher than that achievable via differential equations

and therefore agent-based models might be preferred in biology-related disciplines, despite the lack of analytical results. In other contexts, for reasons of analytical tractability, differential equations are preferred to difference equations, despite the inaccuracy that certain theoretical assumptions might impose on a model, such as that population size is continuous (Colyvan 2013). Similar examples of assumptions introduced for tractability reasons can be found across domains. For instance, classical logic allows that propositions take only two truth values, even if uncertainty and vagueness are standard properties of our statements. In Bayesian epistemology, degrees of belief are represented as probabilities, mainly for the flexibility and formal simplicity of the probability calculus (Hartmann and Sprenger 2010).[4]

Notice that representational robustness must be distinguished from robustness to equivalent or isomorphic mathematical structures. The examples given above illustrate that here we are dealing with models whose mathematical structures describe a certain phenomenon with different degrees of specificity. We are not exploring cases where different mathematical approaches can describe the same objects or equivalent properties of different objects. Representational robustness is thus similar to structural robustness, with the main difference being that, in the former, the variations of interest concern the models' mathematical assumptions, which in turn reflect different aspects of the system under consideration, rather than the variables of a model whose main structure remains fixed.

In a paper on robustness analysis, Kuorikoski et al. (2010) claim that also in economic modeling it is standard to adopt assumptions with the purpose of facilitating the mathematical tractability of a model, even if the simplifications these assumptions introduce do not literally mirror the phenomenon under scrutiny. This a matter of interest for robustness analysis, insofar as it asks whether and under what conditions we should expect that alternative

[4]It might be asked whether the distinction between representational and structural robustness cannot be always traced back to differences in the structure of the initial model. In other words, given that, if a certain assumption is replaced, there are corresponding changes in the parameters of the model, why isn't representational robustness a case of structural robustness? The reason is that structural robustness - rather than investigating the role of different parameters in a model with respect to the target system - focuses on the role of the mathematical assumptions adopted to address a certain problem. It is the purpose of representational robustness to consider the conditions under which we are justified to adopt less accurate assumptions in all those cases where, for reasons of mathematical tractability, we are not in a position to adopt more accurate ones, as the example of the discontinuos functions for continuos cases show.

mathematical assumptions will produce consistent results.

To address this issue, I will follow the analysis provided by Kuorikoski et al. (2010). The case study discussed in their paper refers to a model in economic geography, which is a sub-field of economics that investigates the conditions under which an economic activity agglomerates in a certain region as against the conditions under which it disperses. Since its first formulation by Krugman (1991), the model of this phenomenon, known as the Core-Periphery model, has been analyzed under different assumptions, such as different transport costs and different utility functions.

The purpose of the analysis is to observe which properties break down under different specifications and which ones are robust, even if all the assumptions considered, whether about individuals' utility functions, or transport costs, are unrealistic assumptions. In this respect, robustness analysis makes it possible to observe whether the results of a model strictly depend on one of these false assumptions. If it does not, i.e. if the result is consistent across them, then it might be because a real mechanism has been captured across different formulations: "That the same results obtain with alternative specifications of transportation costs suggests that the results crucially hinge not on the unrealistic assumption of iceberg transportation costs but on the realistic substantial assumption that goods are costly to transport." (Kuorikoski et al. 2010, p. 557).

It might be asked why one should value the consistency of the derivation across modalities if changes in the model's assumptions mainly consist in replacing falsities with other falsities. Even if the authors do not emphasize it, the intuitive answer is that when economists conduct robustness analysis, the goal is to replace certain assumptions with others, which are in a way *less false*, in the sense that they provide a more realistic representation of the real-world phenomenon. Thus, going back to Krugman's example, the fact that transport costs are relevant to the result is not the only aspect that should interest the modelers. It is also important to identify how transport costs change and in function of which variables, since this can affect the interplay of centrifugal and centripetal forces in a relevant sense for the predictions of the Core-Periphery model.

To see whether economists proceed in this way when conducting robustness analysis, let us consider in more detail the case of the transport costs function. The iceberg transport cost function (Samuelson 1952; Krugman 1991) is considered to be one of the major innovations in economic geog-

raphy, whose introduction was crucial to determine a paradigm shift from previous theories of international trade. The function is so called because it is based on the principle that part of the goods melts away when transferred from the place of production to the place of delivery. Even if the iceberg formulation is evidently a theoretical construct, not based on direct observation, still it is considered to be appropriate mainly for two reasons: first, it reflects the idea that goods are costly to transport; secondly, it enables the formulation of transport factors not as a separate component of the model but as part of the goods itself. In the words of Krugman (1998): "In terms of modeling convenience, there turns to be out a spectacular synergy between [...] market structure and 'iceberg' transport costs: not only can one avoid the need to model an additional industry, but because the transport cost between two locations is always a constant fraction of the free-on-board price, the constant elasticity of demand is preserved" (p. 11).

Given that the iceberg cost function is highly idealized, in subsequent formulations of the Core-Periphery model, economic geographers have tried to measure how sensitive the predictions are to that assumption. As has been pointed out, a number of unrealistic features follow from the iceberg cost function McCann (2005). Above all, that the price of the delivered goods increases exponentially with distance, with the implausible consequence that the price of the goods might exceed the value of the goods for larger distances. A further problematic aspect is that the price of the delivered goods increases more rapidly for more expensive goods. Kuorikoski et al. report that in a subsequent study by Ottaviano (2002), the iceberg cost function has been substituted with a linear one, without this invalidating the original result. While a linear function can solve the problem of different growth rates according to different initial prices, it does not similarly solve the one of exponential growth with distance. However, if the latter aspect also affects the dynamics of the Core-Periphery model, then it has to be shown that the predictions of the model are preserved even under functions that do not show the same controversial aspects. I do not claim that it is not possible to come up with such explanation, but that such explanation needs to be given to motivate robustness analysis in a meaningful way. Moreover, this is the aspect that has to be highlighted in order to justify robustness analysis. If not, the method is prone to the criticism, raised by many writers (see Odenbaugh and Alexandrova 2011), that replacing idealized assumptions with other idealized ones does not prove anything relevant with respect to the phenomenon under

study.

The main problem seems to be that the functions that differ from the iceberg one, in not showing the same controversial properties, are difficult to implement in the Core-Periphery model. According to McCann (2005), "It is almost impossible to provide direct comparisons between models with the iceberg assumption and those with other sets of transport costs assumptions embedded in them. [...] This is because these more traditional transport costs functions are analytically incompatible with new economic geography models" (p. 312). One of the reasons why the iceberg cost function was initially introduced was indeed to facilitate the mathematical tractability of a certain problem. If it were possible to adopt a more accurate transport cost function, this would have been done from the beginning.

An analogy from physics, discussed by Hindriks (2006), illustrates this point. In classical mechanics, when Newton first determined the orbit of the planets around the sun, he assumed that there were no interplanetary gravitational forces but only the attraction exerted between the sun and each single planet. Even if this simplification was not justifiable on the basis of the negligible effects of interplanetary attraction, still it was necessary for reasons of mathematical tractability. Only later on, thanks to advancements in mathematics, it became possible to introduce interplanetary attraction into the calculation and to redefine the theory on that basis. This is to say that the replacement of tractability assumptions with new ones might require developments in mathematics that were not immediately available at the time when the model was first formulated.

In economics, the problem is compelling as well, since economic models also tend to have a particularly elaborate formal structure. The way in which a certain assumption is introduced into a model also depends on the role it plays in relation to the other components of the model. The Lotka-Volterra model – where consistent results have been achieved both via differential equations and via agent-based models – is a successful case, but it is also a fortunate one, given that it is fairly uncontroversial how to interpret the results deriving from the different mathematical assumptions at the basis of each approach. Similarly, there are other cases where tractability assumptions have been introduced to approximate solutions to problems that could not be solved analytically. As an example, consider the case of numerical methods in solving differential equations. By contrast, and as the Core-Periphery model shows, it is plausible to expect that, in economics, a more

common situation is one in which – if a certain solution has been derived under a specific tractability assumption – it is not straightforward to relax that assumption and to solve the model under different conditions.

Therefore, even if the role of tractability assumptions is crucial in model building, often it is not innocuous for the final predictions. Ideally, tractability assumptions should be replaced with different ones, which do not have the same controversial properties. Yet, as we have seen, it is not trivial to exchange one assumption with another, especially if they were introduced partly for the purpose of satisfying certain analytical requirements dictated by the formal structure of the model. That is, the difficulties in replacing tractability assumptions with new ones have to do with the reasons why they were initially adopted. In this case, other options have to be considered. One way is to identify – if possible – where the predictions of a certain function diverge from the real-world phenomenon and then to consider only those results that derive under certain regimes. For example, in the case of the transport cost function, given that problems mainly arise for greater distances, only the solutions for short distances should be accepted. Often, however, it is not even possible to quantify the possible errors related to the adoption of certain mathematical simplifications. In these cases, as Newton's example shows, the problem is that of conceiving new mathematical methods that make it possible to eliminate the simplifications of the previous treatments. An alternative possibility is to compare the results of models that rely on different assumptions. However, even this move is not free of difficulties. Above all, if every model is based on assumptions that are only partially plausible, it is not clear which is the model whose predictions should be favored.

To conclude, the main point of this section is that even if in principle robustness analysis is a justifiable method of addressing the issue of tractability assumptions in theoretical models, still it is far from being clear how to proceed in order to conduct it, and especially with models where tractability assumptions are introduced to solve nontrivial analytical problems. Whereas adding variables to a model is a less controversial procedure, replacing tractability assumptions with different ones is not similarly uncomplicated and a number of difficulties emerge when trying to provide a methodological account that regulates this practice.[5]

[5]This claim is not meant to suggest that the validity of robustness analysis has been definitively proved, but that, for that analysis to be a viable option, different models should provide comparable outcomes, which is often not the case in scientific practice. I

5.5 Conclusion

Compare the robustness analysis of scientific models with the analysis conducted in a laboratory experiment. In the latter, in order to test whether the relation under observation is the effect of some possible confounders, scientists try to *disturb* that relation, through changes in the experimental setting. The variations they introduce in the experimental setup represent modifications of the situation in which the event under scrutiny takes place, which might influence its occurrence. In the same spirit, when dealing with theoretical models, scientists analyze whether the predictions are stable despite perturbations in the assumptions of the models. Whereas changes in the experimental setup correspond to variations in the actual mechanism that might affect the phenomenon under consideration, changes in the assumptions of a model intervene on its theoretical structure.

The aim of this chapter was to investigate the notion of robustness analysis in scientific practice and to spell out the details of the procedure. A distinction has been explored between different theoretical assumptions, according to the role they play in the derivations of the model. Following Weisberg's classification (2008), this distinction is between assumptions that concern changes in the system under observation, and assumptions changing the formal description of a phenomenon. In the two cases, scientists are aiming at different goals: when changing the variables of the models, the goal is to observe whether a certain relation is robust to the introduction of other variables. When changing the formal approach, the intention is to assess the impact of the assumptions introduced for reasons of mathematical tractability, since the result of a model should not depend on the specifics of the structure adopted in order to make the derivation possible.

In the paper *Economic Modelling as Robustness Analysis*, (Kuorikoski et al. 2010) claim that robustness analysis plays an essential role in economics and that the method is crucial to increase explanatory and predictive power and to drive progress in their discipline. As evidence for this, they report

suggest, therefore, that it is more urgent to provide a method to deal with those cases where the outcome of different analyses are not easily comparable with one another, rather than debating how to deal with a rather uncommon circumstance. However, I realize that the exploration of one single case study from economic geography, despite its role as a gold standard in the relevant literature, is not conclusive for a more general claim on the status of robustness analysis. Other case studies need to be considered from the literature in support of or against the position defended here.

how it is often enough for economists to prove that the predictions of a model are not robust under modifications of some of the initial assumptions, for example full rationality, to have a publication in a prestigious economic journal. Not surprisingly, criticisms of robustness analysis have been taken seriously as potentially undermining the basis of economic methodology. As shown in the first part of this chapter, a lively debate has ensued in the scientific community in reaction to these critiques.

Overall, I have argued that robustness analysis, in the various senses attributed to it, is in principle a useful method of testing the validity of the predictions of a model. In this respect, if scientific models yield informative predictions, it is also due to robustness analysis, insofar as the method provides a way of securing the predictions of the models. However, several issues have been pointed out that need to be addressed in order to regulate the practice of robustness analysis. Attention has been called in particular to problems related to the robustness analysis of tractability assumptions. I have not focused on the analogous issues of structural robustness and parameter robustness, not because I regard them as insignificant, but mainly because it has been urged by several authors (Cartwright 2005; Kuorikoski et al. 2010) that the impact of tractability assumptions requires a systematic analysis, which is still missing in the philosophy of economics literature.

By means of a case study in economic geography, I have explained that tractability assumptions are adopted chiefly for mathematical reasons, but that their introduction is not necessarily harmless for the predictions of the models. It is only under specific circumstances that the results of a model are not criticizable because they were derived under certain mathematical simplifications. When it is not possible to quantify the errors that tractability assumptions entail, the question of how to proceed in order to evaluate the predictions of a model is far from being straightforward.

The strategy of robustness analysis is to replace tractability assumptions with different ones. However, often for the same reason these assumptions were introduced in the first place, it is difficult to adopt new ones without compromising the overall structure of the model. This is especially the case with economic models that have a complex structure. An alternative strategy is to contrast the results of models relying on different initial assumptions. Yet, there is no obvious reason to expect that the results of different models can be easily compared with each other. Just as different experimental practices might lead to different results, thereby posing the question of how

to interpret these results (Stegenga 2009), the same is true of inconsistent predictions deriving from models with different initial assumptions. Probably, the standard situation in economics is not one where the predictions are stable across conditions, but where different results stem from different analyses, so that one should adopt a cautious attitude to the results. In conclusion, rather than trying to debunk the role of robustness analysis as a general method of assessing the predictions of scientific models, this chapter has tried to highlight the difficulties encountered in the practice of robustness analysis, and to indicate where effective strategies need to be developed in response to these difficulties.

Chapter 6

Conclusions

In conclusion, let us first return to the questions posed at the beginning of this study, summarize the main findings and finally point to directions for further research. The major thread running through this thesis is the emergence of norms in society and norms compliance. To address these issues, a family of formal and experimental methods have been adopted, followed by a methodological reflection on robustness analysis.

More specifically: In **chapter 2**, I have presented an agent-based model of a descriptive norm, which is a behavioral rule that individuals follow when their empirical expectations of others following the same rule are met. An account of the emergence of descriptive norms has been provided by first looking at a simple case, that of the standing ovation. We have then examined the structure of the standing ovation, and showed that it can be generalized to describe the emergence of a wide range of descriptive norms.

Chapter 3 has dealt with a mathematical model for the emergence of descriptive norms, where the individual decision problem is formalized with the standard Bayesian belief revision machinery. Whereas in the previous study the emergence of descriptive norms relied on heuristic modeling, a Bayesian model has provided a more general picture of the emergence of norms and clarified the assumptions made in heuristic models. In this model, the priors formalize the belief that the behavioral rule is a descriptive norm; the evidence is provided by other group members' behavior and the likelihood by their reliability. We have implemented the model in a series of computer simulations and examined the group-level outcomes. We have claimed that domain-general belief revision helps explain why we look for regularities in social life in the first place.

Chapter 4 has addressed the question of how other people's opinion affects judgments of norm transgressions. A modification of the Asch paradigm (1951, 1955) has been adopted to examine conformity in the moral domain. We have asked how peer group opinion alters normative judgments of scenarios involving violations of moral, social, and decency norms. The results have indicated that a norm taxonomy can be based on the insulation of different kinds of norm from other people's beliefs and preferences, even if all kinds of norm are prone to some extent to social pressure.

In **chapter 5**, I have examined the epistemic validity of robustness analysis, as a method of testing whether the predictions of a model are the unintended effect of the initial unrealistic assumptions. In this final chapter, I have argued that even if in principle scientific theories do gain support from robustness analysis, still it is a fortunate case in which the results of different investigations can be compared with one another, and especially if models have a complicated structure and rely on several assumptions. In these cases, a way has to be found to interpret different results as an alternative to suspending judgment. Despite the restriction of its applicability to a limited domain, I have acknowledged robustness analysis as a valuable practice inherent in model building and designing experiments.

Altogether, the projects I have undertaken are part of a broad research program that explores the nature and dynamics of decision-making as mediated by norms. Overall, the mechanisms behind norms compliance involve a large variety of factors, the most important of which are as follows: social representations, social learning, moral reasoning, emotional responses and cultural and anthropological contexts. To address them, contributions from different branches of the humanities and the sciences are needed. It would be naïve even to try to offer an exhaustive account of the subject matter, and this study only covers some facets of the story. The attempt made here has been to address at least some of the key features of the decision-making processes, where individual action, emotions and cognition overlap in a way that might be conducive to the emergence of a new norm.

In this respect, each of the first two chapters of this thesis has provided a general model for the emergence of norms, the one describing the structure of individuals' preferences underlying the decision to comply with a norm, the other the cognitive apparatus that sustains normative behavior. Both are explanatory templates capable of further development into more complex and realistic representations. By virtue of their simplicity and generality, these

models are analogous to the original segregation model by Schelling; and like Schelling's model, they are adequate to subsume a large class of phenomena, but can progressively be enriched with more complex and realistic elements.

In future works, I plan to extend both the heuristic and the rational model to a larger normative domain. In particular, a question that the present work leaves open is how to model normative behaviors that involve higher-level expectations, as in the case of social and moral norms. This leads in turn to the question of how normative expectations emerge in the first place. I consider it important to test the predictions of models not only via computer simulations, but also via laboratory experiments, such as the one presented in the third chapter. The experimental study in that chapter, on conformity and normative judgment, pairs with the two previous models, insofar as it constitutes a first step in the direction of a more fine-grained taxonomy of norms.

It is noteworthy that the formal and experimental approaches I have used throughout this thesis reflect a methodological trend in several areas of philosophy. This way of tackling philosophical problems brings philosophy closer to the sciences. What is standard practice in scientific laboratories, where researchers with diverse skills and expertise collaborate on joint projects, is also being adopted in the philosophical world. Collaborative work in philosophy is essential, both for the division of labour and so that researchers can become acquainted with those practices that are not part of their own background. It is important to bear in mind that this shift has a significant impact on research time. It takes time to learn methods for the conduct of formal and experimental research. Designing experiments, collecting data, re-collecting data to test for possible confounders, writing computer codes and running simulations, are activities that usually span several semesters. These activities lead to the writing of papers, which are in many respects distant from philosophical works in the classical sense, and where the boundaries between philosophy and the sciences become blurred.

Nevertheless, the role of the philosopher in this domain of inquiry is not marginal. This research lies at the intersection of classic philosophical topics such as reasoning, volition and action. Moreover, a study on normative behavior needs to be supported by an ethical analysis that has as its central core the fundamentals of morality, and the distinction between what is right and wrong. In the course of this study, I have mentioned how awareness of the mechanisms of norms compliance might induce pro-social behaviors and

help to dismiss negative norms. To be sure, there are a number of clear-cut cases in which it is evident that the norms under consideration are harmful for the individuals – extreme examples are child marriages or female genital mutilation in African societies. Yet, in many circumstances it is a matter of dispute how to assess whether a norm is positive or negative in content. In this respect, it pertains to the ethicist to guide the debate concerning the evaluative assessment of the norms we live by. We are continually witnessing how dramatic can be those processes of change in norms, where conduct that was considered to be right (or wrong) according to previous standards of behavior, ceases to be such and becomes the expression of new values in society.

To conclude, the studies I have presented in this thesis are the outcome of some of the projects I have undertaken during the years of my PhD at Tilburg University and at the Ludwig Maximilians University Munich. Each of them can be seen as a continuation of the previous one, originally intended to consider the questions left open by the previous study, but afterwards extended to its own domain. Together, they represent the philosophical and experimental journey that has preoccupied me until now, and that began with the intention to learn the trade of formal models, simulations and experiments in order to apply them to a philosophical investigation.

Bibliography

Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, 41: 15–34.

Asch, S. (1951). Effects of group pressure on the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, Leadership and Men*, pp. 177-190, Pittsburg, PA: Carnegie Press.

Asch, S. (1955). Opinions and social pressure. *Scientific American*, 193: 33–35.

Bicchieri, C. (2006). *The Grammar Of Society*. New York, NY: Cambridge University Press.

Bicchieri, C. (2008). The fragility of fairness: An experimental investigation on the conditional status of pro-social norms. *Philosophical Issues*, 18: 229–248.

Bordia, P. (1997). Face-to-face versus computer-mediated communication: A synthesis of the experimental literature. *The Journal of Business Communication*, 34(1): 99–120.

Bovens, L. and Hartmann, S. (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Bruch, E. and Mare, R. (2006). Neighborhood choice and neighborhood change. *American Journal of Sociology*, 112(3): 667–709.

Cartwright, N. (1991). Replicability, reproducibility, and robustness - Comments on Harry Collins. *History of Political Economics*, 23: 143–155.

Cartwright N. (2005). The vanity of rigour in economics: Theoretical models and Galilean experiments. In P. Fontaine and R. J. Leonard (Eds.), *The 'Experiment' in the History of Economics*, pp. 135–153, London: Routledge.

Chater N. and Oaksford, M. (2008). *The Probabilistic Mind*. Oxford: Oxford University Press.

Cialdini, R., Kallgren, C., Reno, R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology*, 24: 201–234.

Cialdini, R., and Goldstein, N. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55: 591–622.

Cinnirella, M., and Green, B. (2007). Does 'cyber-conformity' vary cross-culturally? Exploring the effect of culture and communication medium on social conformity. *Computers in Human Behavior*, 23(4): 2011–2025.

Colyvan, M. (2013). Idealisations in normative models. *Synthese*, 190(8): 1337–1350.

Colyvan, M., and Ginzburg, L.V. (2003). The Galilean turn in population ecology. *Biology and Philosophy*, 18: 401–414.

D'Arms, J., Batterman, R., Górny, K. (1998). Game theoretic explanations and the evolution of justice. *Philosophy of Science*, 65: 76–102.

Elster, J. (2009). Social norms and the explanation of behavior. In P. Hedström and P. Bearman (Eds.), *The Oxford Handbook of Analytical Sociology*, pp. 195–217, Oxford: Oxford University Press.

Frigg, R. and Hartmann, S. (2012). Models in science. *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), ed. by E. Zalta.

Gopnik A. and Tenenbaum J. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Science*, 10(3): 281–287.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1): 3–32.

Greene, J.D., Nystrom, L., Engell, A.D., Darley, J., Cohen, J. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44: 389–400.

Greene, J.D., Sommerville, R., Nystrom, L., Darley, J., Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293: 2105–2108.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences* 14(8): 357–364

Guala, F. (2005). *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Guala, F. and Mittone L. (2005). Experiments in economics: Testing theories vs. the robustness of phenomena. *Journal of Economic Methodology*, 12: 495–515.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge: Cambridge University Press.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108: 814–834.

Haidt, J., Koller, S., Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65: 613–628.

Hartmann, S. and Sprenger, J. (2010). Bayesian Epistemology. In S. Bernecker and D. Pritchard (Eds.), *Routledge Companion to Epistemology*, pp. 609-620, London: Routledge.

Hindriks F. A. (2006). Tractability assumptions and the Musgrave-Mäki typology. *Journal of Economic Methodology*, 13: 401–423.

Hiltz, S.R., Johnson, K., Turoff, M. (1986). Experiments in group decision making communication process and outcome in face-to-face versus computerized conferences. *Human Communication Research*, 13(2): 225–252.

Hogg, M.A., and Reid, S.A. (2006). Social identity, self-categorization, and the communication of group norms. *Communication Theory*, 16(1): 7–30.

Hoover, K. (2006). Fragility and robustness in econometrics: Introduction to the symposium. *Journal of Economic Methodology*, 13(2): 159–160.

Huebner, B., Dwyer, S., and Hauser M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Science*, 13(1): 1–6.

Kelly, D., Stich, S., Haley, K., Eng, S. and Fessler, D. (2007). Harm, affect and the moral/conventional distinction. *Mind and Language*, 22: 117–131.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economics*, 99: 483–499.

Krugman, P. (1998). What's new about the new economic geography? *Oxford Review of Economic Policy*, 14(2): 7–17.

Kuorikoski, J., Lehtinen A., Marchionni C. (2010). Economic modelling as robustness analysis. *British Journal of Philosophy of Science*, 61: 541–567.

Kuorikoski, J., Lehtinen A., Marchionni C. (2012). Robustness analysis disclaimer: Please read the manual before use! *Biology and Philosophy*, 27(6): 891–902.

Jones, M. and Love, B.C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34: 169–231.

Van Lankveld, G., Spronck, P., van den Herik, J., and Arntz, A. (2011). Games as personality profiling tools. In M. Preuss (Ed.), *Proceedings of the 2011 IEEE Conference on Computational Intelligence in Games*, pp. 197–202.

Laporte, L., van Nimwegen, C., Uyttendaele, A.J (2010). Do people say what they think: social conformity behavior in varying degrees of online social presence. In E.B. Hvannberg, M.K. Lárusdóttir, A. Blandford, and J. Gulliksen (Eds.), *Proceedings of NordiCHI 2010*, pp. 305–314.

Leamer, E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73: 31–44.

Lewis, David (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Lisciandra, C. Colombo M., Nilsenova M. (2013). Conformorality. A study of group conditioning of normative judgment. *The Review of Philosophy and Psychology.*

McCann, P. (2005). Transport costs and new economic geography. *Journal of Economic Geography*, 5(3): 305–318.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67: 371–378.

Miller, J. and Page, S.E. (2004). The standing ovation problem. *Complexity*, 9(5): 8–16.

Muldoon, R., Lisciandra, C., Bicchieri, C., Hartmann, S., Sprenger, J. (forthcoming). On the emergence of descriptive norms. *Politics, Philosophy, and Economics.*

Muldoon, R., Lisciandra, C., Hartmann, S. (under review). Why are descriptive norms there?

Muldoon, R., Smith T., Weisberg M. (2012). Segregation that no one seeks. *Philosophy of Science*, 79(1): 38–62.

Nado, J., Kelly, D., Stich, S. (2009). Moral judgment. In J. Symons and P. Calvo (Eds.), *The Routledge Companion to the Philosophy of Psychology*, pp. 621-633, New York, NY: Routledge.

Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, 84: 221–236.

Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment.* New York, NY: Oxford University Press.

Nucci, L. (2001). *Education in the Moral Domain.* Cambridge: Cambridge University Press.

Nucci, L. and Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development*, 49: 400–407.

Odenbaugh, J. and Alexandrova, A. (2011). Buyer beware: robustness analyses in economics and biology. *Biology and Philosophy*, 26: 757–771.

Orzack, S. H. and Sober, E. (1993). A critical assessment of Levins's 'The Strategy of Model Building in Population Biology (1966)'. *The Quarterly Review of Biology*, 68(4): 533–546.

Ottaviano, G., Tabuchi T., Thisse J.F. (2002). Agglomeration and trade revisited. *International Economic Review*, 43(2): 409–436.

Parker, W. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science*, 78(4): 579–600.

Perrin, J. (1923). *Atoms*. Trans D. L. Hammick, New York, NY: Van Nostrand.

Pirtle, Z, Meyer R., Hamilton A. (2010). What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environmental Science and Policy*, 13: 351–161.

Popper, K. (1934). *Logik der Forschung*. Vienna: Springer. Trans. *Logic of Scientific Discovery*. London: Hutchinson, 1959

Prinz, J. (2006). The Emotional Basis of Moral Judgments. *Philosophical Explorations*, 9: 29–43.

Puntoni, S., Langhe, B. de and Osselaer, S. Van (2008). Bilingualism and the emotional intensity of advertising language. *Journal of Consumer Research*, 35: 1012–1025.

Royzman, E., Leeman R., Baron J. (2009). Unsentimental ethics: Towards a content-specific account of the moral-conventional distinction. *Cognition*, 112: 159–174.

Royzman, E., Goodwin, G., Leeman, R. (2011). When sentimental rules collide: Norms with feelings in the dilemmatic context. *Cognition*, 121: 101–114.

Samuelson P. (1952). The transfer problem and transport costs. *Economic Journal*, 64: 264–289.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1: 143–186.

Schelling, T. C. (1978). *Micromotives and Macrobehavior*. New York, NY: W.W. Norton.

Schnall, S., Haidt, J., Clore, G. L., Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 34: 1096–1109.

Short, J., Williams, E., and Christie, B. (1976). *The social psychology or telecommunications*. London: Wiley.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18: 429–434.

Schupbach, J. (2010). Comparing probabilistic measures of explanatory power. *Philosophy of Science*, 78(5): 813–829.

Skyrms, B. (1996). *The Evolution of Social Contract*, Cambridge: Cambridge University Press.

Smetana, J. (1993). Understanding of social rules. In M. Bennett (Ed.) *The Development of Social Cognition: The Child as Psychologist*, pp. 111–141, New York, NY: Guilford Press.

Smilowitz, M., Compton, C., Flint, L. (1988). The effect of computer mediated communication on an individual's judgement: A study based on the methods of Asch's social influence experiment. *Computers in Human Behavior*, 4: 311–321.

Soler, L., Trizio E., Nickles T., Wismatt W. (2012) (Eds). *Characterizing the Robustness of Science*. Boston Studies in the Philosophy of Science, 292

Sousa, P. (2009). On testing the 'moral law'. *Mind and Language*, 24(2): 209–234.

Sousa, P., Holbrook, C., Piazza, J. (2009). The morality of harm. *Cognition*, 113: 80–92.

Stegenga, J. (2009). Robustness, discordance, and relevance. *Philosophy of Science*, 76(5): 650–661.

Stich, S., Fessler, D., Kelly D. (2009). On the morality of harm: A response to Sousa, Holbrook and Piazza. *Cognition*, 113: 93–97.

Sugden R. (1986). *The Economics of Rights, Cooperation and Welfare*. Oxford: Blackwell.

Sugden R. (2001). Credible Worlds: the Status of Theoretical Models in Economics. *Journal of Economic Methodology*, 7(1): 1–31.

Tajfel, H. (1981). *Human Groups and Social Categories*. Cambridge: Cambridge University Press.

Tentori, K. Crupi, V. Bonini N., Osherson D. (2007). Comparison of confirmation measures. *Cognition*, 103(1): 107–119.

Turiel, E. (1977). Distinct conceptual and developmental domains: Social convention and morality. In H. Howe, and C. Keasey. (Eds.), *Nebraska Symposium on Motivation, 1977: Social Cognitive Development, 25*, pp. 77–116, Lincoln, NE: Nebraska University of Press.

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge: Cambridge University Press.

Turiel, E. (2002). *The Culture of Morality: Social Development, Context and Conflict*. Cambridge: Cambridge University Press.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, 73: 730–742.

Weisberg, M (2007). Three kinds of idealizations. *Journal of Philosophy* 104(12): 639–659.

Weisberg, M. and Reisman (2008). The robust Volterra principle. *Philosophy of Science* 75: 106–131.

Wheatley, T. and Haidt, J. (2005). Hypnotically induced disgust makes moral judgments more severe. *Psychological Science*, 16: 780–784.

Whiten, A., Horner, V., de Waal, F. (2005). Conformity to cultural norms of tool use in chimpanzees. *Nature*, 437: 737–740.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, 13(2): 219–40.

Young, H. P. (2009) Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review* 99(5): 1899–1924.

INVITATION

Reception following the defense will take place in the 'Kleine Foyer' at 15.30. Please join us!

Chiara Lisciandra

Finnish Centre of Excellence in the Philosophy of the Social Sciences

University of Helsinki

Department of Political and Economic Studies

PO Box 24, 00014 Helsinki
Finland

C.lisciandra@uvt.nl